

Julio Abascal · Simone Barbosa  
Mirko Fetter · Tom Gross  
Philippe Palanque · Marco Winckler (Eds.)

LNCS 9297

# Human-Computer Interaction – **INTERACT 2015**

15th IFIP TC 13 International Conference  
Bamberg, Germany, September 14–18, 2015  
Proceedings, Part II

2  
Part II



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, Lancaster, UK*

Takeo Kanade

*Carnegie Mellon University, Pittsburgh, PA, USA*

Josef Kittler

*University of Surrey, Guildford, UK*

Jon M. Kleinberg

*Cornell University, Ithaca, NY, USA*

Friedemann Mattern

*ETH Zurich, Zürich, Switzerland*

John C. Mitchell

*Stanford University, Stanford, CA, USA*

Moni Naor

*Weizmann Institute of Science, Rehovot, Israel*

C. Pandu Rangan

*Indian Institute of Technology, Madras, India*

Bernhard Steffen

*TU Dortmund University, Dortmund, Germany*

Demetri Terzopoulos

*University of California, Los Angeles, CA, USA*

Doug Tygar

*University of California, Berkeley, CA, USA*

Gerhard Weikum

*Max Planck Institute for Informatics, Saarbrücken, Germany*

More information about this series at <http://www.springer.com/series/7409>

Julio Abascal · Simone Barbosa  
Mirko Fetter · Tom Gross  
Philippe Palanque · Marco Winckler (Eds.)

# Human-Computer Interaction – INTERACT 2015

15th IFIP TC 13 International Conference  
Bamberg, Germany, September 14–18, 2015  
Proceedings, Part II



Springer

*Editors*

Julio Abascal

Universidad del País Vasco/Euskal Herriko  
Unibertsitatea  
Donostia-San Sebastián  
Spain

Simone Barbosa  
PUC-Rio  
Rio de Janeiro  
Brazil

Mirko Fetter  
University of Bamberg  
Bamberg  
Germany

Tom Gross

University of Bamberg  
Bamberg  
Germany

Philippe Palanque  
University Paul Sabatier  
Toulouse  
France

Marco Winckler  
University Paul Sabatier  
Toulouse  
France

ISSN 0302-9743

Lecture Notes in Computer Science

ISBN 978-3-319-22667-5

DOI 10.1007/978-3-319-22668-2

ISSN 1611-3349 (electronic)

ISBN 978-3-319-22668-2 (eBook)

Library of Congress Control Number: 2015946321

LNCS Sublibrary: SL3 – Information Systems and Applications, incl. Internet/Web, and HCI

Springer Cham Heidelberg New York Dordrecht London

© IFIP International Federation for Information Processing 2015

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

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

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

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media  
([www.springer.com](http://www.springer.com))

## **Foreword**

The 15th IFIP TC.13 International Conference on Human–Computer Interaction, INTERACT 2015, was held during September 14–18, 2015, in Bamberg, Germany, organized by the University of Bamberg. The city of Bamberg is proud of its more than 1,000-year-old center. It has more than 2,400 historically listed buildings and became a UNESCO World Cultural Heritage Site in 1993. With 70,000 inhabitants, Bamberg is a small town in the heart of Europe.

The theme of the 2015 edition was “Connection, tradition, innovation.” In its relatively short history, the human–computer interaction (HCI) area has experienced impressive development. Theories, methodologies, procedures, guidelines, and tools have been progressively proposed, discussed, tested, and frequently adopted by academia and industry. The protagonists of this development created in a short period of time a scientific and technological tradition able to produce high-quality interaction systems. However, the evolution of the computers and networks pose new challenges to all stakeholders. Innovation, based on tradition, is the only way to face these challenges, even if innovation often requires breaking the tradition. In order to make this process possible, INTERACT 2015 provides diverse and abundant connection opportunities. A multidisciplinary approach is characteristic of the HCI field. INTERACT 2015 aimed to connect all the matters able to contribute to the quality of the future interactions among people and computers.

The series of INTERACT international conferences (started in 1984) is supported by Technical Committee 13 on Human–Computer Interaction of the International Federation for Information Processing (IFIP). This committee aims at developing the science and technology of the interaction between humans and computing devices.

IFIP was created in 1960 under the auspices of UNESCO with the aim of balancing worldwide the development of computer technology and Science. Technical Committee 13 is fully conscious of the social importance of information and communication technologies for our world, today and in the future. Therefore, INTERACT 2015 made efforts to attract and host people from all over the world, and to pay attention to the constraints imposed on HCI by differences in culture, language, technological availability, physical, as well as sensory and cognitive differences, among other dimensions of interest.

INTERACT 2015 gathered a stimulating collection of research papers and reports of development and practice that acknowledge the diverse disciplines, abilities, cultures, and societies, and that address all the aspects of HCI, including technical, human, social, and esthetic.

Like its predecessors, INTERACT 2015 aimed to be an exciting forum for communication with people of similar interests, to foster collaboration and learning. Being by nature a multidisciplinary field, HCI requires interaction and discussion among diverse people with different interests and backgrounds. INTERACT 2015 was directed both to the academic and industrial world, always highlighting the latest developments

in the discipline of HCI and its current applications. Experienced HCI researchers and professionals, as well as newcomers to the HCI field, interested in the design or evaluation of interactive software, development of new technologies for interaction, and research on general theories of HCI met in Bamberg.

We thank all the authors who chose INTERACT 2015 as the venue to publish their research. This was again an outstanding year for the conference in terms of submissions in all the technical categories.

We received 651 submissions. Of these, the following were accepted: 93 full research papers; 74 short research papers; eight demos; 30 interactive posters; four organizational overviews; three panels; six tutorials; 11 workshops; and 13 doctoral consortium papers.

The acceptance rate for the full papers was 29.6 % and 26.8 % for short papers.

In order to select the highest-quality contributions, an elaborate review system was organized including shepherding of 38 full research papers that went through a second and sometimes a third round of review. That process was primarily handled by the 32 meta-reviewers who willingly assisted and ensured the selection of high-quality full research papers to be presented at INTERACT 2015.

The final decision on acceptance or rejection of papers was taken in a plenary Program Committee meeting held in Tampere (Finland) in February 2015, aimed to discuss a consistent set of criteria to deal with inevitable differences among the large number of reviewers who were recruited and supported by the meta-reviewers. The technical program chairs and the track chairs, the general chairs, and the members of IFIP Technical Committee 13 participated in the meeting.

Special thanks must go to the track chairs and all the reviewers, who put in an enormous amount of work to ensure that quality criteria were maintained throughout the selection process. We also want to acknowledge the excellent work of the co-chairs of the different sections of the conference and the meta-reviewers of the full research paper track.

We also thank the members of the Organizing Committee, especially Mirko Fetter, local organization chair, who provided us with all the necessary resources to facilitate our work. Finally, we wish to express a special thank you to the proceedings publication chair, Marco Winckler, who did extraordinary work to put this volume together.

September 2015

Tom Gross  
Julio Abascal  
Simone Barbosa  
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 of 37 national societies and nine 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. From 2017 the conference series will become an annual conference.

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 worldwide.

The main orientation is toward users, especially 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; modeling 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 multimedia systems to match system needs and organizational practices, etc.).

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. 2007 IFIP TC 13 also launched an accessibility award to recognize an outstanding contribution with international impact in the field of accessibility for disabled users in HCI. In 2013, IFIP TC 13 launched the Interaction Design for International Development (IDID) Award, which recognizes the most outstanding contribution to the application of interactive systems for social and economic development of people in

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

IFIP TC 13 also recognizes pioneers in the area of HCI. An IFIP TC 13 pioneer is one who, through active participation in IFIP Technical Committees or related IFIP groups, has made outstanding contributions to the educational, theoretical, technical, commercial, or professional aspects of analysis, design, construction, evaluation, and use of interactive systems. IFIP TC 13 pioneers are appointed annually and awards are handed over at the INTERACT conference.

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 here.

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.

WG 13.5 (Resilience, Reliability, Safety, and Human Error in 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 will promote the use of knowledge, concepts, methods, and techniques that enable 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) aims to establish a study and research program that will combine both scientific work and practical applications in the fields of HCI and visualization. It will integrate 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.

WG13.8 (Interaction Design and International Development) are currently working to reformulate their aims and scope.

WG13.9 (Interaction Design and Children) aims to support practitioners, regulators, and researchers to develop the study of interaction design and children across international contexts.

New Working Groups are formed as areas of significance to HCI arise. Further information is available on the IFIP TC13 website: <http://ifip-tc13.org/>

## IFIP TC13 Members

### Officers

**Chair**

Jan Gulliksen, Sweden

**Vice-chair**

Philippe Palanque, France

**Vice-Chair for WG and SIG**

Simone D.J. Barbosa, Brazil

**Treasurer**

Anirudha Joshi, India

**Secretary**

Marco Winckler, France

**Webmaster**

Helen Petrie, UK

### Country Representatives

**Australia**

Henry B.L. Duh  
Australian Computer Society

**Austria**

Geraldine Fitzpatrick  
Austrian Computer Society

**Belgium**

Monique Noirhomme-Fraiture  
Fédération des Associations  
Informatiques de Belgique

**Brazil**

Raquel Oliveira Prates  
Brazilian Computer Society (SBC)

**Bulgaria**

Kamelia Stefanova  
Bulgarian Academy of Sciences

**Canada**

Heather O'Brien  
Canadian Information Processing Society

**Chile**

Jaime Sánchez  
Chilean Society of Computer Science

**Croatia**

Andrina Granic  
Croatian Information Technology  
Association (CITA)

**Cyprus**

Panayiotis Zaphiris  
Cyprus Computer Society

**Czech Republic**

Zdeněk Míkovec  
Czech Society for Cybernetics &  
Informatics

**Denmark**

Torkil Clemmensen  
Danish Federation for Information  
Processing

**Finland**

Kari-Jouko Räihä  
Finnish Information Processing  
Association

**France**

Philippe Palanque  
Société des Electriciens et des  
Électroniciens (SEE)

**Germany**

Tom Gross  
Gesellschaft für Informatik

**Hungary**

Cecilia Sik Lanyi  
John V. Neumann Computer  
Society

**Iceland**

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

**India**

Anirudha Joshi  
Computer Society of India

**Ireland**

Liam J. Bannon  
Irish Computer Society

**Italy**

Fabio Paternò  
Italian Computer Society

**Japan**

Yoshifumi Kitamura  
Information Processing Society of Japan

**Korea**

Gerry Kim  
KIISE

**Malaysia**

Chui Yin Wong  
Malaysian National Computer  
Confederation

**The Netherlands**

Vanessa Evers  
Nederlands Genootschap voor  
Informatica

**New Zealand**

Mark Apperley  
New Zealand Computer Society

**Nigeria**

Chris C. Nwannenna  
Nigeria Computer Society

**Norway**

Dag Svanes  
Norwegian Computer Society

**Poland**

Marcin Sikorski  
Poland Academy of Sciences

**Portugal**

Pedro Campos  
Associação Portuguesa para o Desen-  
volvimento da Sociedade da Infor-  
mação (APDSI)

**Slovakia**

Vanda Benešová  
The Slovak Society for Computer  
Science

**South Africa**

Janet L. Wesson  
The Computer Society of South Africa

**Spain**

Julio Abascal  
Asociación de Técnicos de Informática  
(ATI)

**Sweden**

Jan Gulliksen  
Swedish Computer Society

**Switzerland**

Solange Ghernaouti  
Swiss Federation for Information  
Processing

**Tunisia**

Mona Laroussi  
Ecole Supérieure des Communications  
De Tunis (SUP'COM)

**UK**

Andy Dearden  
British Computer Society (BCS)

**USA**

Gerrit van der Veer  
Association for Computing Machinery  
(ACM)

## Expert Members

Nikos Avouris (Greece)  
Simone D.J. Barbosa (Brazil)  
Peter Forbrig (Germany)  
Joaquim Jorge (Portugal)  
Paula Kotzé (South Africa)  
Masaaki Kurosu (Japan)

Gitte Lindgaard (Australia)  
Zhengjie Liu (China)  
Fernando Loizides (Cyprus)  
Dan Orwa (Kenya)  
Frank Vetere (Australia)

## Working Group Chairs

### **WG13.1 (Education in HCI and HCI Curricula)**

Konrad Baumann, Austria

### **WG13.2 (Methodologies for User-Centered System Design)**

Marco Winckler, France

### **WG13.3 (HCI and Disability)**

Helen Petrie, UK

### **WG13.4 (also 2.7) (User Interface Engineering)**

Jürgen Ziegler, Germany

### **WG13.5 (Resilience, Reliability, Safety and Human Error in System Development)**

Chris Johnson, UK

### **WG13.6 (Human–Work Interaction Design)**

Pedro Campos, Portugal

### **WG13.7 (HCI and Visualization)**

Achim Ebert, Germany

### **WG 13.8 (Interaction Design and International Development)**

José Adbelnour Nocera, UK

### **WG 13.9 (Interaction Design and Children)**

Janet Read, UK

## **Conference Organizing Committee**

### **General Conference Co-chairs**

Tom Gross, Germany

Julio Abascal, Spain

### **Tutorials Co-chairs**

Christoph Beckmann, Germany

Regina Bernhaupt, France

### **Full Papers Chairs**

Simone D.J. Barbosa, Brazil

Philippe Palanque, France

### **Workshops Co-chairs**

Christoph Beckmann, Germany

Víctor López-Jaquero, Spain

### **Short Papers Co-chairs**

Fabio Paternò, Italy

Kari-Jouko Räihä, Finland

### **Doctoral Consortium Co-chairs**

Geraldine Fitzpatrick, Austria

Panayiotis Zaphiris, Cyprus

### **Posters and Demos Co-chairs**

Stephen Brewster, UK

David McGookin, UK

### **Proceedings Chair**

Marco Winckler, France

### **Organization Overviews Co-chairs**

Melanie Fitzgerald, USA

Kori Inkpen, USA

### **Madness Co-chairs**

Artur Lugmayr, Finland

Björn Stockleben, Germany

Tim Merritt, Denmark

### **Panels Co-chairs**

Anirudha N. Joshi, India

Gitte Lindgaard, Australia

### **Local Organization Co-chairs**

Mirko Fetter, Germany

Claudia Tischler, Germany

### **Open Space Co-chairs**

Christoph Beckmann, Germany

Achim Ebert, Germany

### **Student Volunteers Co-chairs**

Robert Beaton, USA

Sascha Herr, Germany

## **Program Committee**

### **Meta-reviewers**

Birgit Bomsdorf, Germany

Gaëlle Calvary, France

José Campos, Portugal

Pedro Campos, Portugal

Luca Chittaro, Italy

Torkil Clemmensen, Denmark

Paul Curzon, UK

Achim Ebert, Germany

Peter Forbrig, Germany

Michael Harrison, UK

Anirudha Joshi, India  
Denis Lalanne, Switzerland  
Effie Law, UK  
Célia Martinie, France  
Laurence Nigay, France  
Monique Noirhomme, Belgium  
Fabio Paternò, Italy  
Helen Petrie, UK  
Antonio Piccinno, Italy  
Aaron Quigley, UK  
Kari-Jouko Räihä, Finland  
Virpi Roto, Finland

Luciana Salgado Cardoso de Castro,  
Brazil  
Paula Alexandra Silva, Ireland  
Frank Steinicke, Germany  
Simone Stumpf, UK  
Allistair Sutcliffe, UK  
Jean Vanderdonckt, Belgium  
Gerhard Weber, Germany  
Astrid Weiss, Austria  
Marco Winckler, France  
Panayiotis Zaphiris, Cyprus

## Reviewers

José Abdelnour-Nocera, UK  
Al Mahmud Abdullah, Australia  
Silvia Abrahão, Spain  
Funmi Adebesin, South Africa  
Ana Paula Afonso, Portugal  
David Ahlström, Austria  
Pierre Akiki, Lebanon  
Deepak Akkil, Finland  
Hannu Alen, Finland  
Jan Alexandersson, Germany  
José Carlos Bacelar Almeida, Portugal  
Florian Alt, Germany  
Julian Alvarez, France  
Junia Coutinho Anacleto, Brazil  
Leonardo Angelini, Switzerland  
Craig Anslow, New Zealand  
Mark Apperley, New Zealand  
Nathalie Aquino, Paraguay  
Liliana Ardissono, Italy  
Carmelo Ardito, Italy  
Oscar Javier Ariza Núñez, Germany  
Myriam Arrue, Spain  
Ilhan Aslan, Austria  
Simon Attfield, UK  
Nikolaos Avouris, Greece  
Chris Baber, UK  
Myroslav Bachynskyi, Germany  
Jonathan Back, UK  
Gilles Bailly, France  
Liam Bannon, Ireland

Emilia Barakova, The Netherlands  
Javier Barcenila, France  
Louise Barkhuus, USA  
Barbara Rita Barricelli, Italy  
Valentina Bartalesi, Italy  
Mohammed Basher, Saudi Arabia  
Christoph Beckmann, Germany  
Yacine Bellik, France  
Vanda Benešová, Slovak Republic  
Kawtar Benghazi, Spain  
David Benyon, UK  
François Bérard, France  
Regina Bernhaupt, Austria  
Karsten Berns, Germany  
Nadia Berthouze, UK  
Raymond Bertram, Finland  
Mark Billinghamurst, New Zealand  
Dorrit Billman, USA  
Silvia Amelia Bim, Brazil  
Fernando Birra, Portugal  
Renaud Blanch, France  
Ann Blandford, UK  
Mads Boedker, Denmark  
Davide Bolchini, USA  
Birgit Bomsdorf, Germany  
Rodrigo Bonacin, Brazil  
Paolo Gaspare Bottoni, Italy  
Fatma Bouali, France  
Chris Bowers, UK  
Giorgio Brajnik, Italy

- Anke Brock, France  
Barry Brown, Sweden  
Judith Brown, Canada  
Gerd Bruder, Germany  
Duncan Brumby, UK  
Nick Bryan-Kinns, UK  
Stéphanie Buisine, France  
Sabin-Corneliu Buraga, Romania  
Paris Buttfield-Addison, Australia  
Maria Claudia Buzzi, Italy  
Marina Buzzi, Italy  
Cristina Cachero, Spain  
Sybille Caffiau, France  
Paul Cairns, UK  
Roberto Caldara, Switzerland  
Gaëlle Calvary, France  
Licia Calvi, The Netherlands  
José Campos, Portugal  
Pedro Campos, Portugal  
Katia Canepa Vega, Brazil  
Maria-Dolores Cano, Spain  
Maria Beatriz Carmo, Portugal  
Francesco Carrino, Switzerland  
Stefano Carrino, Switzerland  
Luis Carriço, Portugal  
Marcus Carter, Australia  
Daniel Cernea, Germany  
Teresa Chambel, Portugal  
Stéphane Chatty, France  
Monchu Chen, Portugal  
Yu Chen, Switzerland  
Kelvin Cheng, Singapore  
Yoram Chisik, Portugal  
Luca Chittaro, Italy  
Elizabeth Churchill, USA  
Torkil Clemmensen, Denmark  
Gilbert Cockton, UK  
Karin Coninx, Belgium  
Tayana Conte, Brazil  
Stéphane Conversy, France  
Jeremy Cooperstock, Canada  
Nuno Correia, Portugal  
Joëlle Coutaz, France  
Céline Coutrix, France  
Nadine Couture, France  
Chris Creed, UK  
Martin Cronel, France  
James Crowley, France  
Jácome Cunha, Portugal  
Paul Curzon, UK  
Marie d'Udekem, Belgium  
Florian Daiber, Germany  
Girish Dalvi, India  
José Danado, UK  
Antonella De Angeli, Italy  
Alexander De Luca, Switzerland  
Maria De Marsico, Italy  
Giorgio De Michelis, Italy  
Leonardo Cunha de Miranda, Brazil  
Boris De Ruyter, The Netherlands  
Clarisse de Souza, Brazil  
Alexandre Demeure, France  
Giuseppe Desolda, Italy  
Ines Di Loreto, France  
Paulo Dias, Portugal  
Shalaka Dighe, India  
Christian Dindler, Denmark  
Anke Dittmar, Germany  
Pierre Dragicevic, France  
Carlos Duarte, Portugal  
Cathy Dudek, Canada  
Henry Been-Lirn Duh, Australia  
Bruno Dumas, Belgium  
Sophie Dupuy-Chessa, France  
Achim Ebert, Germany  
Florian Echtler, Germany  
Rob Edlin-White, UK  
Jan Engelen, Belgium  
Thomas Erickson, USA  
Elina Eriksson, Sweden  
Dominik Ertl, UK  
Parisa Eslambolchilar, UK  
Marc Fabri, UK  
Carla Faria Leitão, Brazil  
Ava Fatah gen Schieck, UK  
Xavier Ferre, Spain  
Eija Ferreira, Finland  
Mirko Fetter, Germany  
Sebastian Feuerstack, Germany  
Vagner Figueiredo de Santana, Brazil  
Daniela Fogli, Italy  
Joan Fons, Spain

Manuel Fonseca, Portugal	Martin Hitz, Austria
Peter Forbrig, Germany	Thuong Hoang, Australia
Marcus Foth, Australia	Rüdiger Hoffmann, Germany
Andre Freire, Brazil	Jennifer Horkoff, UK
Carla D.S. Freitas, Brazil	Heiko Hornung, Brazil
Jonas Fritsch, Denmark	Ko-Hsun Huang, Taiwan, Republic of China
Luca Frosini, Italy	Alina Huldtgren, The Netherlands
Dominic Furniss, UK	Ebba Thora Hvamberg, Iceland
Nestor Garay-Vitoria, Spain	Aulikki Hyrskykari, Finland
Jérémie Garcia, France	Ioanna Iacovides, UK
Roberto García, Spain	Netta Iivari, Finland
Jose Luis Garrido, Spain	Mirja Ilves, Finland
Franca Garzotto, Italy	Yavuz İnal, Turkey
Isabela Gasparini, Brazil	Poika Isokoski, Finland
Miguel Gea, Spain	Minna Isomursu, Finland
Patrick Gebhard, Germany	Howell Istance, Finland
Cristina Gena, Italy	Ido A. Jurgel, Germany
Giuseppe Ghiani, Italy	Mikkel R. Jakobsen, Denmark
Patrick Girard, France	Francis Jambon, France
Kentaro Go, Japan	Jacek Jankowski, Poland
Daniel Gonçalves, Portugal	Maddy Janse, The Netherlands
Rúben Gouveia, Portugal	Nuno Jardim Nunes, Portugal
Nicholas Graham, Canada	Caroline Jay, UK
Andrina Granic, Croatia	Kasper Løvborg Jensen, Denmark
Toni Granollers, Spain	Mikael Johnson, Finland
Saul Greenberg, Canada	Matt Jones, UK
John Grundy, Australia	Joaquim Jorge, Portugal
Nuno Guimaraes, Portugal	Rui Jose, Portugal
Jan Gulliksen, Sweden	Anirudha Joshi, India
Rebecca Gulotta, USA	Christophe Jouffrais, France
Mieke Haesen, Belgium	Anne Joutsenvirta, Finland
Hans Hagen, Germany	Marko Jurmu, Finland
Jonna Häkkilä, Finland	Eija Kaasinen, Finland
Jukka Häkkinen, Finland	Jari Kangas, Finland
Jaakko Hakulinen, Finland	Anne Marie Kanstrup, Denmark
Lynne Hall, UK	Victor Kaptelinin, Sweden
Arnaud Hamon, France	Evangelos Karapanos, Portugal
Chris Harrison, USA	Kristiina Karvonen, Finland
Daniel Harrison, UK	Dinesh Katre, India
Michael Harrison, UK	Manolya Kavakli, Australia
Ruediger Heimgaertner, Germany	Patrick Gage Kelley, USA
Tomi Heimonen, Finland	Ryan Kelly, UK
Matthias Heintz, UK	Rabia Khan, UK
Ingi Helgason, UK	Hideki Koike, Japan
Susan Catherine Herring, USA	Christophe Kolski, France
Wilko Heuten, Germany	

- Hannu Korhonen, Finland  
Nataliya Kosmyna, France  
Paula Kotze, South Africa  
Christian Kray, Germany  
Per Ola Kristensson, UK  
Sari Kujala, Finland  
Todd Kulesza, USA  
Denis Lalanne, Switzerland  
David Lamas, Estonia  
Michael Lankes, Austria  
Rosa Lanzilotti, Italy  
Przemyslaw Lasota, USA  
Yann Laurillau, France  
Effie Law, UK  
Shaimaa Lazem, UK  
Xavier Le Pallec, France  
Eric Lecolinet, France  
Jong-Seok Lee, South Korea  
Asko Lehmuskallio, Finland  
Antti Leino, Finland  
Juha Leino, Finland  
Tuomas Leisti, Finland  
Jair Leite, Brazil  
Alexander Lenz, UK  
Barbara Leporini, Italy  
Sophie Lepreux, France  
Karen Y. Li, UK  
Edirlei Lima, Brazil  
James Lin, USA  
Mats Lind, Sweden  
Agnes Lisowska Masson, Switzerland  
Zhengjie Liu, China  
Sara Ljungblad, Sweden  
Corrado lo Storto, Italy  
Steffen Lohmann, Germany  
Fernando Loizides, Cyprus  
Víctor López-Jaquero, Spain  
Fabien Lotte, France  
Maria Dolores Lozano, Spain  
Yichen Lu, Finland  
Paul Lubos, Germany  
Stephanie Ludi, USA  
Bernd Ludwig, Germany  
Andreas Luedtke, Germany  
Christopher Lueg, Australia  
Jo Lumsden, UK  
Christof Lutteroth, New Zealand  
Kris Luyten, Belgium  
Anderson Maciel, Brazil  
I. Scott MacKenzie, Canada  
Allan MacLean, UK  
Christian Maertin, Germany  
Charlotte Magnusson, Sweden  
Ana Gabriela Maguitman, Argentina  
Päivi Majaranta, Finland  
Marco Manca, Italy  
Nicolai Marquardt, UK  
Célia Martinie, France  
Paolo Masci, UK  
Masood Masoodian, New Zealand  
Maristella Matera, Italy  
Denys J.C. Matthies, Germany  
Peter W. McOwan, UK  
Gerrit Meixner, Germany  
Guy Melançon, France  
Amaia Mendez Zorrilla, Spain  
Maria Menendez Blanco, Italy  
Zdenek Mikovec, Czech Republic  
Jan-Torsten Milde, Germany  
Nicole Mirnig, Austria  
Giulio Mori, Italy  
Roxana Morosanu, UK  
Christiane Moser, Austria  
Marcelle Mota, Brazil  
Omar Mubin, Australia  
Chrystie Myketiak, UK  
Miguel Nacenta, UK  
Lennart Nacke, Canada  
Mathieu Nancel, Canada  
Bonnie Nardi, USA  
David Navarre, France  
Ather Nawaz, Norway  
Luciana Nedel, Brazil  
Alexandra Nemery, France  
Vania Neris, Brazil  
Daniel Nesbitt, UK  
Lene Nielsen, Denmark  
Anton Nijholt, The Netherlands  
Laurence Nigay, France  
Manuel Noguera, Spain  
Monique Noirhomme, Belgium  
Julianne Nyhan, UK  
Clemens Nylandsted Klokmose,  
Denmark

- Michael O Grady, Ireland  
Aisling Ann O’Kane, UK  
Marianna Obrecht, UK  
Lars Oestreicher, Sweden  
Jarno Ojala, Finland  
Patrick Oladimeji, UK  
Kathia Oliveira, France  
Thomas Olsson, Finland  
Dan Orwa, Kenya  
Nuno Otero, Sweden  
Benoit Otjacques, Luxembourg  
Saila Ovaska, Finland  
Janne Paavilainen, Finland  
Xinru Page, USA  
Ana Paiva, Portugal  
Jose Ignacio Panach Navarrete, Spain  
Eleftherios Papachristos, Greece  
Konstantinos Papoutsakis, Greece  
Avi Parush, Israel  
Oscar Pastor, Spain  
Fabio Paternò, Italy  
Celeste Lyn Paul, USA  
Andriy Pavlovych, Canada  
Roberto Pereira, UK  
Vinícius Carvalho Pereira, Brazil  
Mark J. Perry, UK  
Hele Petrie, UK  
Antoinio Piccinno, Italy  
Lara Piccolo, UK  
Emmanuel Pietriga, France  
Thomas Pietrzak, France  
Frank Pollick, UK  
Ravi Poovaiah, India  
Roman Popp, Austria  
Christopher Power, UK  
Raquel Prates, USA  
Costin Pribeanu, Romania  
Angel Puerta, USA  
Kai Puolamäki, Finland  
Victor M.R. Penichet, Spain  
Aaron Quigley, UK  
Kari-Jouko Räihä, Finland  
Roope Raisamo, Finland  
Venkatesh Rajamanickam, India  
Nitendra Rajput, India  
Ismo Rakkolainen, Finland  
Jussi Rantala, Finland  
Alberto Raposo, Brazil  
Dimitrios Raptis, Denmark  
Umar Rashid, UK  
Kirsten Rassmus-Gröhn, Sweden  
Matthias Rauterberg, The Netherlands  
Janet Read, UK  
Mandryk Regan Lee, Canada  
Patrick Reignier, France  
Christian Remy, Switzerland  
Karen Renaud, UK  
Yann Riche, USA  
Fabien Ringeval, Germany  
Thomas Rist, Germany  
Paola Rodriguez, Colombia  
Markus Rohde, Germany  
Teresa Romão, Portugal  
Jose Rouillard, France  
Virpi Roto, Finland  
Thijs Roumen, Germany  
Gustavo Alberto Rovelo Ruiz, Belgium  
Elisa Rubegni, Switzerland  
Simon Ruffieux, Switzerland  
Jaime Ruiz, USA  
Angel Ruiz-Zafra, Spain  
Rimvydas Ruksenas, UK  
Horacio Saggion, Spain  
Pascal Salembier, France  
Luciana Salgado Cardoso de Castro,  
Brazil  
Antti Salovaara, Finland  
Leonardo Sandoval, UK  
Carmen Santoro, Italy  
Corina Sas, UK  
Andreas Savva, UK  
Taufique Sayeed, Austria  
Gianluca Schiavo, Italy  
Antonio Giovanni Schiavone, Italy  
Albrecht Schmidt, Germany  
Stefan Schneegass, Germany  
Kevin Schneider, Canada  
Vinicius Segura, Brazil  
Marcos Serrano, France  
Ehud Sharlin, Canada  
Sumita Sharma, Finland  
Moushumi Sharmin, USA  
Abhishek Shrivastava, India  
Beat Signer, Belgium

Harri Siirtola, Finland	Susan Ellen Turner, UK
Paula A. Silva, Ireland	Markku Turunen, Finland
Bruno S. Silva, Brazil	Blase Ur, USA
Carlos CL Silva, Portugal	Heli Väätäjä, Finland
João Carlos Silva, Portugal	Stefano Valtolina, Italy
Jose Luis Silva, Portugal	Judy van Biljon, South Africa
Paula Alexandra Silva, Ireland	Jos P. van Leeuwen, The Netherlands
Milene Silveira, Brazil	Paul van Schaik, UK
Carla Simone, Italy	Jeroen Vanattenhoven, Belgium
Shamus Smith, Australia	Jean Vanderdonckt, Belgium
Andreas Sonderegger, Switzerland	Jari Varsaluoma, Finland
Keyur Sorathia, India	Radu-Daniel Vatavu, Romania
Fabio Sorrentino, Italy	Angel Velazquez-Iturbide, Spain
Hamit Soyle, UK	Hanna Venesvirta, Finland
Oleg Spakov, Finland	Jayant Venkatanathan, India
Lucio Davide Spano, Italy	Gilles Venturini, France
Mark Vincent Springett, UK	Arnold Vermeeren, The Netherlands
Jan Stage, Denmark	Karel Vermeulen, UK
Christian Stary, Austria	Frédéric Vernier, France
Katarzyna Stawarz, UK	Markel Vigo, UK
Frank Steinicke, Germany	Nadine Vigouroux, France
Gerald Stollnberger, Austria	Chris Vincent, UK
Markus Stolze, Switzerland	Giuliana Vitiello, Italy
Simone Stumpf, UK	Arnd Vitzthum, Germany
Noi Sukaviriya, USA	Dhaval Vyas, Australia
Allistar Sutcliffe, UK	Mike Wald, UK
David Mark Swallow, UK	Jim Wallace, Canada
Tapio Takala, Finland	Tanja Carita Walsh, Finland
Chee-wee Tan, Denmark	Robert Walter, Germany
Franck Tarpin-Bernard, France	Leon Watts, UK
Carlos Teixeira, Portugal	Gerhard Weber, Germany
Luis Teixeira, Portugal	Rina Wehbe, Canada
Daniel Tetteroo, The Netherlands	Astrid Weiss, Austria
Jakob Tholander, Sweden	Janet Louise Wesson, South Africa
Nigel Thomas, UK	Graham Wilson, UK
Liisa Tiittula, Finland	Stephanie Wilson, UK
Nava Tintarev, UK	Marco Winckler, France
Martin Tomitsch, Australia	Theophilus Winschiers, Namibia
Ilaria Torre, Italy	Chui Yin Wong, Malaysia
Marilyn Tremaine, USA	Wolfgang Wörndl, Germany
Daniela Trevisan, Brazil	Volker Wulf, Germany
Sanjay Tripathi, India	Yeliz Yesilada, Turkey
Janice Tsai, USA	Salu Ylirisku, Finland
Manfred Tscheligi, Austria	Nur Haryani Zakaria, Malaysia
Huawei Tu, UK	Massimo Zancanaro, Italy
Outi Tuisku, Finland	Panayiotis Zaphiris, Cyprus
Phil Turner, UK	Jürgen Ziegler, Germany

## Sponsors and Supporters

### Sponsors



### Supporters



## Contents – Part II

### Computer-Supported Cooperative Work and Social Computing

EmbodiNet: Enriching Distributed Musical Collaboration Through Embodied Interactions . . . . .	1
<i>Dalia El-Shimy and Jeremy R. Cooperstock</i>	
Preference Elicitation and Negotiation in a Group Recommender System . . . . .	20
<i>Jesús Omar Álvarez Márquez and Jürgen Ziegler</i>	
The #selfiestation: Design and Use of a Kiosk for Taking Selfies in the Enterprise . . . . .	38
<i>Casey Dugan, Sven Laumer, Thomas Erickson, Wendy Kellogg, and Werner Geyer</i>	
The LuminUs: Providing Musicians with Visual Feedback on the Gaze and Body Motion of Their Co-performers . . . . .	47
<i>Evan Morgan, Hatice Gunes, and Nick Bryan-Kinns</i>	
An Artifact Ecology in a Nutshell: A Distributed Cognition Perspective for Collaboration and Coordination . . . . .	55
<i>Christina Vasilou, Andri Ioannou, and Panayiotis Zaphiris</i>	
Assessing a Collaborative Application for Comic Strips Composition . . . . .	73
<i>Eleonora Mencarini, Gianluca Schiavo, Alessandro Cappelletti, Oliviero Stock, and Massimo Zancanaro</i>	
Augmenting Collaborative MOOC Video Viewing with Synchronized Textbook . . . . .	81
<i>Nan Li, Łukasz Kidziński, and Pierre Dillenbourg</i>	
EXCITE: EXploring Collaborative Interaction in Tracked Environments . . . . .	89
<i>Nicolai Marquardt, Frederico Schardong, and Anthony Tang</i>	
The Usefulness of Method-Resources for Evaluating a Collaborative Training Simulator . . . . .	98
<i>Ebba Thora Hvannberg, Gyda Halldorsdottir, and Jan Rudinsky</i>	

### End-User Development

Flat Design vs Traditional Design: Comparative Experimental Study . . . . .	106
<i>Ivan Burmistrov, Tatiana Zlokazova, Anna Izmalkova, and Anna Leonova</i>	

How to Organize the Annotation Systems in Human-Computer Environment: Study, Classification and Observations . . . . .	115
<i>Anis Kalboussi, Nizar Omheni, Omar Mazhoud, and Ahmed Hadj Kacem</i>	
Mini-Orb: A Personal Indoor Climate Preference Feedback Interface . . . . .	134
<i>Markus Rittenbruch, Jared Donovan, and Yasuhiro Santo</i>	
Prototyping the Self-Authored Video Interview: Challenges and Opportunities . . . . .	150
<i>Stephen Snow, Markus Rittenbruch, and Margot Brereton</i>	
<b>Evaluation Methods/Usability Evaluation</b>	
An Empirical Study of the Effects of Three Think-Aloud Protocols on Identification of Usability Problems . . . . .	159
<i>Anders Bruun and Jan Stage</i>	
An Observational Study of How Experienced Programmers Annotate Program Code . . . . .	177
<i>Craig J. Sutherland, Andrew Luxton-Reilly, and Beryl Plimmer</i>	
Around-Device Interactions: A Usability Study of Frame Markers in Acquisition Tasks . . . . .	195
<i>Fernando Garcia-Sanjuan, Alejandro Catala, Geraldine Fitzpatrick, and Javier Jaen</i>	
On Applying Experience Sampling Method to A/B Testing of Mobile Applications: A Case Study . . . . .	203
<i>Myunghee Lee and Gerard J. Kim</i>	
Usability Aspects of the Inside-in Approach for Ancillary Search Tasks on the Web . . . . .	211
<i>Marco Winckler, Ricardo Cava, Eric Barboni, Philippe Palanque, and Carla Freitas</i>	
Using Affinity Diagrams to Evaluate Interactive Prototypes . . . . .	231
<i>Andrés Lucero</i>	
What Users Prefer and Why: A User Study on Effective Presentation Styles of Opinion Summarization . . . . .	249
<i>Xiaojun Yuan, Ning Sa, Grace Begany, and Huahai Yang</i>	
A Comparison of Five HSV Color Selection Interfaces for Mobile Painting Search . . . . .	265
<i>Min Zhang, Guoping Qiu, Natasha Alechina, and Sarah Atkinson</i>	

Computer-Related Attribution Styles: Typology and Data Collection Methods . . . . .	274
<i>Adelka Niels and Monique Janneck</i>	
Reciprocity in Rapid Ethnography: Giving Back by Making the Small Things Count . . . . .	292
<i>Pieter Duysburgh and Karin Slegers</i>	
Testing the Unknown – Value of Usability Testing for Complex Professional Systems Development . . . . .	300
<i>Kimmo Tarkkanen, Ville Harkke, and Pekka Reijonen</i>	
<b>Eye Tracking</b>	
An Empirical Investigation of Gaze Selection in Mid-Air Gestural 3D Manipulation . . . . .	315
<i>Eduardo Velloso, Jayson Turner, Jason Alexander, Andreas Bulling, and Hans Gellersen</i>	
Four Eyes See More Than Two: Shared Gaze in the Car . . . . .	331
<i>Sandra Trösterer, Magdalena Gärtner, Martin Wuchse, Bernhard Maurer, Axel Baumgartner, Alexander Meschtscherjakov, and Manfred Tscheligi</i>	
Gaze+touch vs. Touch: What's the Trade-off When Using Gaze to Extend Touch to Remote Displays? . . . . .	349
<i>Ken Pfeuffer, Jason Alexander, and Hans Gellersen</i>	
<b>Gesture Interaction</b>	
Gestu-Wan - An Intelligible Mid-Air Gesture Guidance System for Walk-up-and-Use Displays . . . . .	368
<i>Gustavo Rovelo, Donald Degraen, Davy Vanacken, Kris Luyten, and Karin Coninx</i>	
Natural Interaction with Video Environments Using Gestures and a Mirror Image Avatar . . . . .	387
<i>Christian Kray, Dennis Wilhelm, Thore Fechner, and Morin Ostkamp</i>	
Sci-Fi Gestures Catalog: Understanding the Future of Gestural Interaction . . . . .	395
<i>Lucas S. Figueiredo, Mariana Pinheiro, Edvar Vilar Neto, Thiago Chaves, and Veronica Teichrieb</i>	
TV Interaction Beyond the Button Press: Exploring the Implications of Gesture, Pressure and Breath as Interaction Mechanisms for a TV User Interface . . . . .	412
<i>Regina Bernhardt, Antoine Desnos, Michael Pirker, and Daniel Schwaiger</i>	

**HCI and Security**

“I Agree”: The Effects of Embedding Terms of Service Key Points in Online User Registration Form . . . . .	420
<i>Matjaž Kljun, Jernej Vičič, Klen Čopić Puciha, and Branko Kavšek</i>	
Automatic Privacy Classification of Personal Photos . . . . .	428
<i>Daniel Buschek, Moritz Bader, Emanuel von Zezschwitz, and Alexander De Luca</i>	
CipherCard: A Token-Based Approach Against Camera-Based Shoulder Surfing Attacks on Common Touchscreen Devices . . . . .	436
<i>Teddy Seyed, Xing-Dong Yang, Anthony Tang, Saul Greenberg, Jiawei Gu, Bin Zhu, and Xiang Cao</i>	
Digital Signage Effectiveness in Retail Stores . . . . .	455
<i>Mari Ervasti, Juha Häikiö, Minna Isomursu, Pekka Isomursu, and Tiina Liuska</i>	
Toward a Deeper Understanding of Data Analysis, Sensemaking, and Signature Discovery . . . . .	463
<i>Sheriff Jolaoso, Russ Burtner, and Alex Endert</i>	

**HCI for Developing Regions and Social Development**

HCI Practices in the Nigerian Software Industry . . . . .	479
<i>Abiodun Ogunyemi, David Lamas, Emmanuel Rotimi Adagunodo, and Isaias Barreto da Rosa</i>	
Penan’s Oroo’ Short Message Signs (PO-SMS): Co-design of a Digital Jungle Sign Language Application . . . . .	489
<i>Tariq Zaman and Heike Winschiers-Theophilus</i>	
The Whodunit Challenge: Mobilizing the Crowd in India. . . . .	505
<i>Aditya Vashistha, Rajan Vaish, Edward Cutrell, and William Thies</i>	
Wayfinding Behavior in India. . . . .	522
<i>Naveed Ahmed</i>	

**HCI for Education**

Evaluating Digital Tabletop Collaborative Writing in the Classroom . . . . .	531
<i>Philip Heslop, Anne Preston, Ahmed Kharrufa, Madeline Balaam, David Leat, and Patrick Olivier</i>	

Evaluating the Accuracy of Pre-kindergarten Children Multi-touch Interaction . . . . .	549
<i>Vicente Nacher and Javier Jaen</i>	
The 5-Step Plan: Empowered Children’s Robotic Product Ideas . . . . .	557
<i>Lara Lammer, Astrid Weiss, and Markus Vincze</i>	
Using IMUs to Identify Supervisors on Touch Devices . . . . .	565
<i>Ahmed Kharrufa, James Nicholson, Paul Dunphy, Steve Hodges, Pam Briggs, and Patrick Olivier</i>	
Design and Usability Evaluation of Adaptive e-learning Systems Based on Learner Knowledge and Learning Style . . . . .	584
<i>Mohammad Alshammary, Rachid Anane, and Robert J. Hendley</i>	
How Does HCI Research Affect Education Programs? A Study in the Brazilian Context . . . . .	592
<i>Isabela Gasparini, Simone Diniz Junqueira Barbosa, Milene Selbach Silveira, Sílvia Amélia Bim, and Clodis Boscaroli</i>	
MindMiner: A Mixed-Initiative Interface for Interactive Distance Metric Learning . . . . .	611
<i>Xiangmin Fan, Youming Liu, Nan Cao, Jason Hong, and Jingtao Wang</i>	
<b>Author Index . . . . .</b>	<b>629</b>

# EmbodiNet: Enriching Distributed Musical Collaboration Through Embodied Interactions

Dalia El-Shimy<sup>(✉)</sup> and Jeremy R. Cooperstock

Centre for Interdisciplinary Research in Music, Media and Technology,  
McGill University, Montreal, QC, Canada  
`{dalia,jer}@cim.mcgill.ca`

**Abstract.** This paper presents EmbodiNet, a novel system that augments distributed performance with dynamic, real-time, hands-free control over several aspects of the musicians' sound, enabling them to seamlessly change volume, affect reverb and adjust their mix. Musical performance is a demanding activity necessitating multiple levels of communication among its participants, as well as a certain degree of creativity, playfulness and spontaneity. As a result, distributed musical performance presents a challenging application area for the “same time/different place” category of Computer-Supported Cooperative Work (CSCW). In fact, musicians wishing to play together over a network are typically limited by tools that differ little from standard videoconferencing. Instead, we propose leveraging the technology inherent to the distributed context towards meaningfully augmenting collaborative performance. In order to do so without introducing new paradigms that may require learning or that may distract musicians from their primary task, we have designed and evaluated embodied controls that capitalize on existing interpersonal interactions. Further designed to restore the spatial properties of sound that are typically absent in the distributed context, and apply the notion of “shared space” found in CSCW research, EmbodiNet also helps confer a greater level of co-presence than standard distributed performance systems. This paper describes the implementation of EmbodiNet, along with the results of a long-term collaboration and experiment with a three-piece band. The long-term collaboration helped illustrate the benefits of augmenting an artistic form of distributed collaboration, and resulted in a system that not only helped enhance our users’ sense of enjoyment and self-expression, but one that they would also likely use in the future.

## 1 Introduction

According to Ackerman, one of the challenges central to the field of Computer-Supported Cooperative Work can be described as the “social-technical gap”, a mismatch resulting from the flexible and nuanced nature of human activity when contrasted with the rigid and brittle nature of technical systems [1]. Thus, the author continues, bridging this gap through computational entities (e.g., information transfer, roles, and policies) that are also flexible and nuanced in nature,

is essential to the successful design of CSCW applications. This is particularly crucial for distributed collaborative environments, where participants often suffer from a lowered sense of shared awareness, and a decrease in mutual perception of non-verbal cues (e.g., gaze direction, gestures, posture) [38]. Such a problem has been tackled extensively by conventional investigations of videoconferencing technologies: telepresence systems, shared virtual table environments (SVTEs) and mobile remote presence (MRP) systems have all emerged in a bid to enrich social engagement within the distributed context. However, such systems strive to improve collaborations of a functional nature, or cooperation on specific, work-related tasks among remote participants [21]. In an effort to explore the breadth of human activity that computer-mediated communication can enrich, we became particularly interested in examining the creative, ludic and spontaneous aspects of social interaction within a distributed context. An additional motivation was exploring whether distributed collaboration could improve on its co-present counterpart by leveraging its underlying technology towards further assisting target users in effectively accomplishing the activity at hand. One area particularly suited for such investigations, given its socially and temporally exacting nature, is that of distributed musical performance.

We decided to examine such challenges by taking a user-driven approach to the design of an augmented distributed performance environment. By choosing an application area where communication is strongly driven by creativity, self-expression and spontaneity, we wanted to explore the ways we could better support the “highly flexible, nuanced, and contextualized” aspects of human activity [1]. Furthermore, as Corness and Schiphorst explain, “[p]erformers tacitly know how to pay close attention to bodily cues that accompany movement, as they have consciously developed their awareness of these cues to enable skilled interaction with other performers” [15]. Thus, we hoped that capitalizing on embodied performer-performer interactions would offer the added advantage of enabling musicians to use our system’s functionality without detaching themselves from the higher level task of performance. Finally, by creating a system that allows musicians to experiment with paradigms that traditional performance does not offer, we sought to examine whether the distributed version of a collaborative activity could offer unique benefits of its own.

Our efforts resulted in EmbodiNet, an augmented distributed performance environment that allows musicians to utilize common gestures and behaviours, such as head tilting, body turning and simple motion, as a means of affecting each other’s volume and reverb levels, adjusting audio mixes and experiencing spatialized sound. EmbodiNet was designed for relaxed performance settings that include room for improvisation or experimentation (e.g., loose rehearsals or jams). An example use case scenario for our system would involve geographically displaced friends who wish to play music together over a network, but seek alternatives to traditional videoconferencing that can further enrich their interpersonal interactions. EmbodiNet can currently only support electric or electronic (rather than acoustic) instruments, in order to ensure that the modified audio mix played back through the musicians’ headphones is not overshadowed by the actual sound of their instruments.

To the best of our knowledge, EmbodiNet is the only distributed performance system of its kind that simultaneously: (1) exports the notion of “shared space” from the CSCW domain to the distributed performance context, allowing musicians to perceive local and remote environments as simple extensions of one another, (2) uses shared space as a means to restore the spatialization of musical instruments that is inherent to the co-present context, yet lost in the distributed one, (3) capitalizes on embodied interactions as a means of control, and (4) offers performers the ability to affect one another’s sound parameters through their interpersonal interactions. Together, such properties allow EmbodiNet to confer a greater level of co-presence than traditional solutions for online performance. The results of our long-term study with a three-piece band confirm that musicians found EmbodiNet to be enjoyable and useful, and that they would likely use it again in the future.

## 2 Related Works

Given its interdisciplinary nature, our work draws inspiration from a variety of research areas. While existing systems for distributed performance have naturally influenced our work [13, 14, 26], we were interested in performance environments developed specifically to explore the implications of the network as “a space for being” [34], rather than simply mimic co-present performance. Examples include Barbosa’s Public Sound Project [3], Tanaka and Bongers’ Global String [39], Braasch et al.’s ViMic system [8], as well as the works of Rebello, Schroeder and Renaud, which emphasize the network as both an acoustic and social medium [31, 32, 35]. A thorough overview of distributed performance environments in relation to our work is presented elsewhere in references [20].

It should come as no surprise, however, that the act of distributing performance over a network would have a strong impact on the nature and level of communication between remote musicians. Renaud, for instance, explains that “[i]nteraction is a real issue in network performance systems as natural visual or sensory cues, such as breathing and gesture, are completely removed from context” [9]. To this, Kapur adds that “[w]aiting backstage to go on, and important aspects of socialisation after a performance, are not the same over a network”, leading to a “loss of society within the band” [26]. In that sense, distributed performance shares a common challenge with Computer-Supported Cooperative Work, a set of activities that also often exhibit a decreased sense of mutual awareness and spontaneous interaction [16, 38]. In fact, we regard distributed performance as a unique application area of the “same time/different place” category of CSCW [25]. As such, the design of systems for distributed performance can benefit from a number of CSCW research topics aiming to facilitate remote collaboration. One such area relevant to our work is that of awareness, described by Dourish and Bly as the ability to know “who is ‘around’, what activities are occurring, who is talking with whom” [16]. Awareness entails a certain level of transparency among remote participants, allowing them to develop a sense of trust and community that, in turn, encourages the playful and creative

sides of interaction that are crucial to successful musical collaboration. Another concern is providing support for the “rich set of social behaviours and cues that we as humans know and share” [2], such as body postures, subtle movements, gaze direction, room acoustics, joint interactions, eye contact and other forms of non-verbal communication [17]. As Sirkin and Ju explain, “[w]e use embodied non-verbal communications such as gestures, body movements, posture, visual orientation, and spatial behavior in concert with our verbal communication to signal our attention, express emotions, convey attitudes, and encourage turn-taking, and... we (perhaps subconsciously) prefer that our technological counterparts follow suit” [38]. An example of a system designed to support such cues, and which directly inspired our display topology, is Hydra, a set of independent communication units, each with its own video display, microphone and speaker. As such, when distributed on a local participant’s desk, Hydra units allow for the spatial and acoustical separation of remote collaborators [37]. Another example from CSCW research that came to influence our work is Ishii’s notion of shared workspaces, conceived as continuous extensions of individual work areas that afford a seamless, two-way transition between collaborative and individual modes of work. In fact, we argue that shared workspaces, as seen in the TeamWorkStation [23] and ClearBoard projects [24], exemplify the philosophy found in the literature on distributed performance of being “in” the network, and described above. Such ideas led to our design of a system configuration that can support the illusion of “shared space”, as described later in this paper. The relationship between CSCW and distributed performance is further expounded upon elsewhere in reference [19].

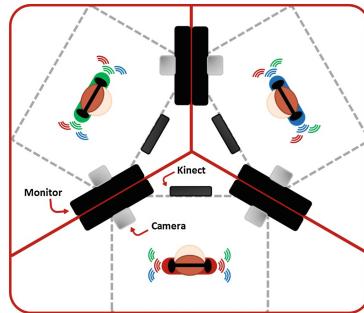
Finally, Computer-Supported Cooperative Work research illustrates that successful collaboration over a network is contingent not only on resolving technological challenges, but also on the development of interaction paradigms that can support both the complexities and subtleties of cooperative behaviour [1, 7, 11, 12, 33]. However, exporting this notion to musical context begets some interesting implications: musical performance is a temporally exacting activity, demanding multiple levels of communication between the players [14]. Therefore, any novel interfaces aiming to augment or facilitate such an activity should be designed, whenever possible, with the intent of reducing the cognitive load they may pose on their users, and avoid distracting them from the higher-level task of performance. As such, embodied interaction, a notion built on the premise of capitalizing on a “broader spectrum of human skills and abilities” [28], lends itself quite naturally to the design of musical interfaces. Our work exemplifies the definition of embodied interaction provided by Antle et al., who describe such an approach as “leveraging users’ natural body movement in direct interaction with spaces and everyday objects to control computational systems” [5]. In fact, embodied interaction has proven to be a suitable option for the design of many non-utilitarian applications [4, 29], including musical interfaces [5, 6, 15].

In summary, to situate our work within the research areas described above, EmbodiNet is a distributed performance environment that offers musicians the illusion of “shared space”, and allows them to utilize embodied interactions to manipulate sound parameters, with the aim of augmenting and improving a

unique form of online collaboration. The ideas behind EmbodiNet reside at the intersection of CSCW and distributed performance research, two fields we believe share many similar challenges and yet which, with very few exceptions [21], have yet to benefit from a full bidirectional flow of information.

### 3 System Description

Our performance environment was deployed across three separate locations. In order to provide musicians with the illusion that they could physically interact in relation to one another, we used a “shared space” metaphor, whereby each of the musicians’ local spaces are mapped onto the Cartesian plane such that they border one another without overlapping, creating, in essence, one large seamless area. Such a configuration for three musicians can be seen in Fig. 1. This solution, in turn, allows the *virtual locations* of remote musicians to appear as though located within an extension of each local musician’s space. When applied to a scenario with three musicians, the virtual locations of remote collaborators places them on either side of the local musician. To support this configuration, every location was equipped with two monitors, each displaying a view of one of the remote spaces. To prevent users from falling out of view as they move about their space, ensure reasonable support for eye contact and, in turn, confer a greater sense of mutual awareness, a camera was mounted behind each monitor, thereby maintaining a line of view between the distributed musicians. Tracking of user position and orientation was carried out with Microsoft Kinect units.



**Fig. 1.** Mapping of three musician locations to create a sense of shared space.

#### 3.1 Features

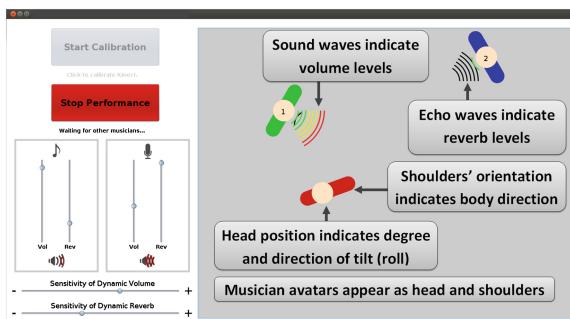
The current implementation of EmbodiNet encompasses five unique features:

- **Dynamic Volume:** As one musician moves *towards* or *away* from another’s virtual location, both can experience each other’s instrument sounds as gradually increasing or decreasing in volume.
- **Dynamic Reverb:** As one musician moves *away* or *toward* another’s virtual location, both can experience each other’s instrument sounds as gradually increasing or decreasing in reverberation, or “reverb”.

- **Mix Control:** A local musician can change the mix of his instrument with those of the remote musicians by tilting his head in the direction where he wants to concentrate the sound of his own instrument. The remote instruments continue to be heard in either left or right headphones as appropriate to their direction.
- **Track Panning:** A local musician can isolate each of the tracks of the remote musicians by changing his body’s orientation.
- **Musician Spatialization:** A local musician can experience the remote musicians’ instruments as spatialized sound sources within his own, local space.

### 3.2 Graphical User Interface

EmbodiNet supplements shared video with a simple graphical user interface (GUI), seen in Fig. 2, appearing on a computer monitor positioned in front of each musician. Not only does the GUI give the musicians complete control over the system features, it also provides simple yet effective dynamic visual representations of the state of their performance at a glance, in an effort to further increase their level of mutual awareness.



**Fig. 2.** Graphical user interface, which includes a control panel and animated graphics. The local musician’s avatar is in red (Color figure online).

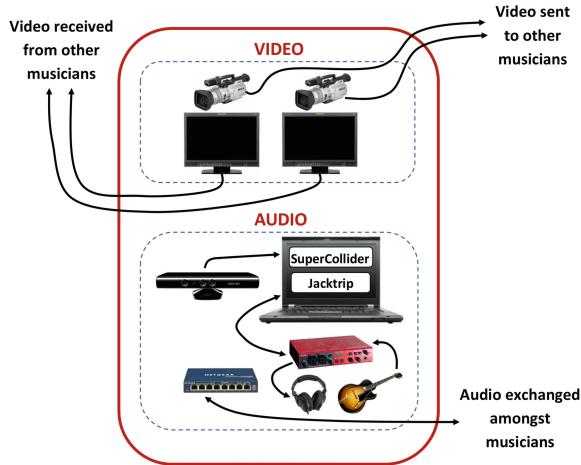
### 3.3 Configuration

EmbodiNet’s hardware configuration can be seen in Fig. 3. We opted to create a simple yet stable setup using analog cameras connected directly to Panasonic BT-LH1700W production monitors that were, in turn, located on either side of each computer monitor displaying EmbodiNet’s GUI. As described earlier, each location includes two monitors, with a camera mounted behind each to maintain a reasonable line of sight across the distributed musicians. Each musician’s instrument, along with a microphone for verbal communication and singing, are plugged into a Roland Edirol FA-101, an audio capture interface. The signals are then processed through SuperCollider, an open source environment and programming language for real-time audio synthesis and algorithmic composition,

where they are adjusted in accordance with the system features described above. The audio streams from SuperCollider are subsequently shared among all three locations through Jacktrip, a tool designed for multi-machine network performance over the Internet. To further reduce delay and guarantee sound stability, a real-time kernel is used on all machines executing Jacktrip, and a Local Area Network (LAN) was created to connect them through a Netgear ProSafe 8 Port Gigabit Switch. Finally, each musician is able to hear his own individual mix through a pair of Sennheiser HD Pro 280 closed headphones.

We measured the end-to-end latency between locations resulting from our hardware and software configurations to be approximately 16 ms. Although Carôt et al. have argued that the maximum delay tolerated by musicians can depend on an ensemble’s style, performance speed and rhythm [10], the “Ensemble Performance Threshold” of 25 ms is commonly regarded as a value beyond which distributed musicians begin to experience difficulties remaining in sync [36]. As such, we note that, while much research in network musical performance has focused on decreasing latency to levels that musicians could tolerate, our aim was not to replicate such results, but rather design and evaluate interactions that could further augment distributed performance environments where latency can be considered a non-issue.

The musicians’ position and orientation data was captured using a Microsoft Kinect, and sent to our SuperCollider software via OpenSoundControl (OSC) messages.



**Fig. 3.** Hardware configuration for EmbodiNet.

## 4 Long-Term Collaboration

EmbodiNet evolved through a series of prototypes and formal user tests, a process depicted in Fig. 4. In the interest of space, we will not detail our previous

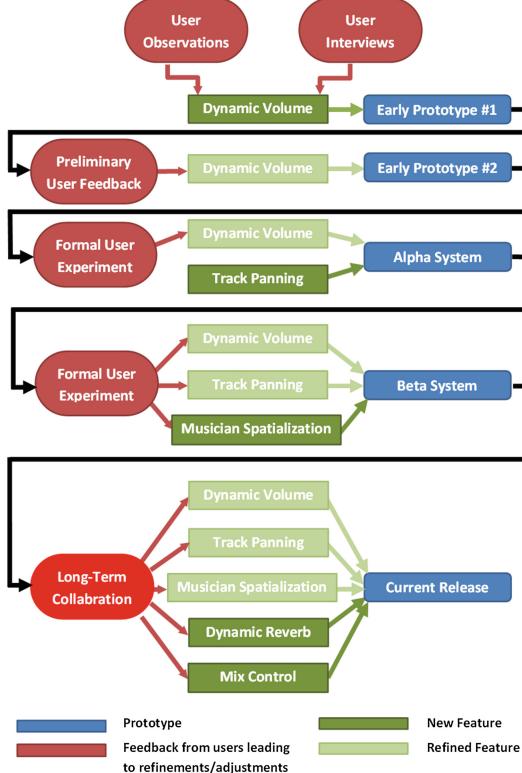
efforts here, although they are described elsewhere in references [18, 20]. In this section, we instead describe the a long-term collaboration and experiment (highlighted in red in Fig. 4) that led to the current implementation of EmbodiNet described in this paper.

Throughout the evolution of EmbodiNet, we had noted that the ‘one-off’ nature of traditional formal user experiments did not provide us with an opportunity to test the effects of small, iterative changes to our system on a regular basis. Furthermore, we questioned whether feedback from first-time users might be biased by the novelty of the system. Thus, after we implemented the beta version of EmbodiNet, which, as seen in Fig. 4 had come to include the Dynamic Volume, Track Panning and Musician Spatialization features, we were motivated to elicit feedback that went beyond simple novelty effects and initial impressions. Inspired by Grudin’s views on the importance of long-term system evaluations within CSCW research [22], and the success of such a methodology within the contexts of both remote collaboration [27] and musical performance [15], we sought to combine the benefits of quantifiable, repeatable user studies and the rich feedback inherent to participatory design by merging elements of both methodologies into a long-term testing and collaboration cycle.

As a result, we invited a band consisting of a 25-year-old guitarist, a 26-year-old keyboardist—both of whom also alternated lead and backup vocals—and a 22-year-old bassist for a series of performance sessions with our beta system. All three were male, and had performed together approximately once per week for almost two years. An introductory brainstorming session was first held, allowing us to showcase our existing system features to the band members, and discuss our vision for the long-term collaboration. Subsequently, we organized weekly meetings that combined formal, quantitative tests with informal, yet in-depth, qualitative discussions.

Since our goal was to “to discover rather than to verify” the effects of each system feature on these various aspects of performance, we knew that the qualitative experiment framework proposed by Ravasio et al. was most suitable to our needs [30]. Therefore, we employed both of their techniques of separation/segmentation and adjection/intensification to design a number of sessions, each focusing on a different feature of the system through an A/B/A-style test, where musicians performed once without the feature, once with the feature, then once again without the feature. At the beginning of each session, musicians were asked to select their base volume levels collaboratively until they reached a satisfactory mix. It is those base levels that our features would subsequently affect during condition B. Each condition lasted approximately 15–20 min, or the time it took the musicians to play through three songs. Musicians were not required to carry out any specific tasks under each condition, only to perform songs of their choice, while voicing to one another or to the test instructor any feelings or concerns they may have throughout the session.

The participants also completed post-condition questionnaires tailored to assess three performance criteria that our previous work, as described elsewhere in Ref. [18], had shown musicians in general tend to deem valuable. Namely, these were enjoyment, creativity and self-expression. Position and orientation data was



**Fig. 4.** User-Centered evolution of EmbodiNet (Color figure online).

collected throughout, along with video footage and audio recordings. After the formal test component of each session, an open discussion in the style of a non-leading interview was held. Musicians were loosely probed about their approach towards the performance and their feelings about the system, and encouraged to provide criticisms, along with suggestions for improvement.

#### 4.1 Session 1: Musician Spatialization

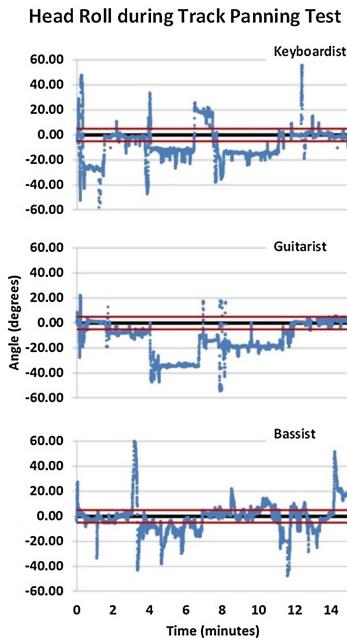
Our first session with the band was designed to focus on the Musician Spatialization feature, whereby the sounds of remote instruments are perceived as emanating from the correct spatial position within the musician's local environment. This feature helps mimic the spatialization effects naturally experienced in a co-present setting, where performers can easily perceive the distance and direction of other instruments surrounding them based on their position and orientation. In that manner, Musician Spatialization was designed to restore some of the natural acoustic dynamics that are lost in typical distributed performance.

Unlike other system features, no explicit gesture is required to activate Musician Spatialization: as long as the feature is enabled, audio from remote musicians will continue to be rendered. However, our post-test discussion with the musicians revealed that the “passive” nature of the feature had somewhat confused them. The guitarist, for instance, explained:

*“I could tell there were changes happening when there were changes happening, but I really had difficulty at times making sense of it.”*

Although its mapping was discussed with them before the performance, the musicians continued to look for a “triggering” gesture that would allow them to control the effect. After additional explanation regarding the feature’s passive nature was offered, the musicians reflected further on their performance, and subsequently indicated that they would be inclined to try it again in light of their new understanding. We suspect that the explanation of the feature we had originally provided lacked sufficient clarity, seeing as the very same implementation of Musician Spatialization was eventually met with more success when the musicians were given another opportunity to test it in Session 5, described below.

Analysis of position and orientation data did not reveal any significant changes in behaviour when Musician Spatialization was used.



**Fig. 5.** Head roll, or tilting, data for all three musicians when Track Panning was used. The red line represents the threshold of  $+/-5$  degrees, beyond which the feature was activated (Color figure online).

## 4.2 Session 2: Track Panning

The second session focused on the Track Panning feature, again through the form of an A/B/A test, followed by a discussion. At the time of testing, Track Panning had been implemented as a function of head roll, and the Mix Control feature had not yet been conceived of. As seen in Fig. 5, orientation data from the formal tests indicates that, while all three musicians experimented with the feature, the keyboardist and guitarist felt more inclined to sustain their interaction for longer periods of time. The guitarist, in particular, regularly isolated the bass track by turning his head to the left, and explained later that it helped him maintain his rhythm.

During the post-test discussion, some of the musicians criticized the head-tilting gesture of the Track Panning feature, noting that it would feel more “natural” to turn one’s *body*, rather than tilt one’s head, towards the virtual location of another musician on whose track they wanted to focus.

Nonetheless, the musicians did appreciate the practical aspect of the function. For instance, when asked to envision use case scenarios for such a feature during performance, the keyboardist explained:

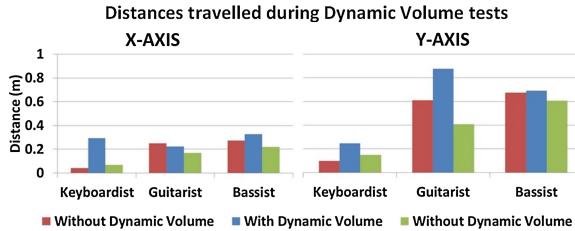
*“Well, mid-performance, say there was a part in the song where a few people were harmonizing together, if I could turn to the screen and we could hear each other better that way, like that would be practical for sure.”*

The musicians suggested that the head tilting gesture would be better suited to listening closely to a mix, as musicians often do in a studio setting, leaning their heads into one headphone at a time. This gave rise to the idea behind the Mix Control feature, whereby a local musician could listen to his own instrument gradually being mixed with either of the remote musicians’ one at a time, simply by tilting his head in the direction corresponding to the remote musician’s virtual location. Musicians had an opportunity to test this new feature, along with an updated version of Track Panning in Session 5, as described below.

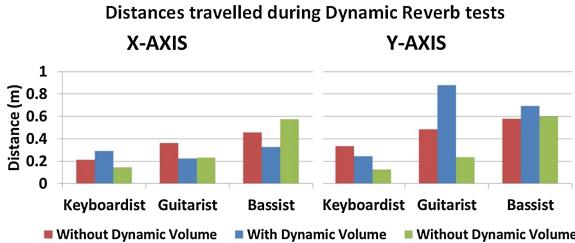
## 4.3 Session 3: Dynamic Volume

The third session included an A/B/A test of the Dynamic Volume feature. Analysis of position data, shown in Fig. 6, revealed that the use of this feature generally helped encourage all three musicians to increase the range of space they covered, rather than maintaining a fixed location, as they were inclined to do otherwise.

During the post-test discussion, the musicians also expressed their interest in controlling another aspect of their sound beyond volume level, namely reverberation. This was considered as a suitable addition to enhance creativity, allowing the musicians to experiment with different sounds. According to the musicians, an increase in reverb when moving further away from each other’s virtual locations could further enhance their feeling of shared space, giving them a more concrete sense of dimension due to the “echoing” nature of this effect.



**Fig. 6.** Distances travelled by all musicians during dynamic volume tests.



**Fig. 7.** Distances travelled by all musicians during dynamic reverb tests.

#### 4.4 Session 4: Dynamic Reverb

We held an interim session where the musicians were invited to experiment with reverb used to simulate rooms of different sizes, and help design the overall effect. Subsequently, the fourth session was centered on the A/B/A testing of the newly implemented “Dynamic Reverb” feature.

Similar to the earlier Dynamic Volume feature, Dynamic Reverb helped increase the interpersonal interaction between musicians, and generally encouraged them to take full advantage of the available space (see Fig. 7). Furthermore, in the post-test discussion, the musicians revealed that they were quite pleased with the feature, with the guitarist stating:

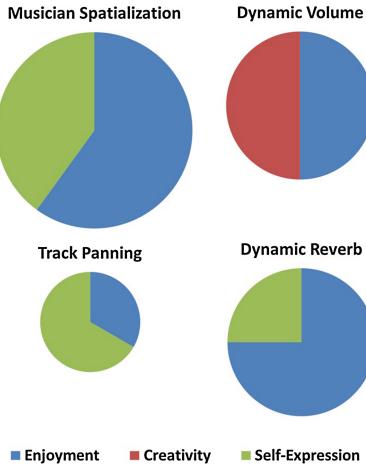
*“I felt that it kind of reacted how I would have wanted it to. It felt a bit like I was able to use it and predict how it was gonna be. It was cool.”*

Furthermore, the guitarist added that while this feature did not necessarily serve a utilitarian purpose, it had an overall positive impact on the performance’s aesthetics:

*“I thought it sounded great, like I just liked the sound. A bit of wetness... It doesn’t really have so much utility so much as it is just an aesthetic thing, it feels natural to have it on”*

#### 4.5 Session 5: Freestyle

The fifth session was a “freestyle” performance: the musicians were simply asked to jam for an hour, selecting which features to turn on or off throughout according to their needs. This session also provided the musicians with the opportunity



**Fig. 8.** Effect of each feature on performance criteria. The size of each pie represents the total number of times, across all criteria, that an improvement was marked.

to test the newly-implemented Mix Control feature, as well as the new versions of Track Panning, now a function of body orientation rather than head tilt. The performance was again followed by a discussion, where musicians provided their opinions of the overall state our system had reached as a result of our on-going collaboration.

Having had the chance to re-visit Musician Spatialization in light of their improved understanding of its functionality, the feature proved to be popular with the guitarist and keyboardist, who were able to finely control it, now that the mapping had been made clearer to them. When asked whether they would use the system in a scenario where they could not be physically co-located, all three agreed that the features would be quite beneficial in facilitating distributed collaboration. The keyboardist, for instance, stated:

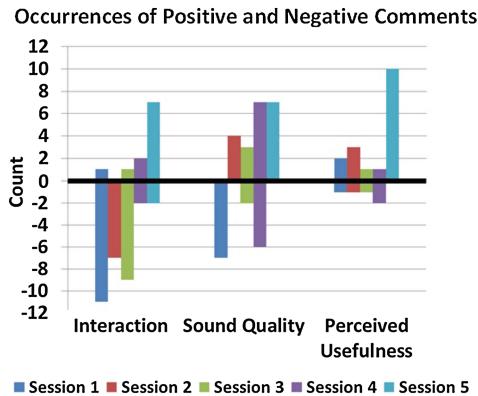
*"I think it's like, if we're doing something like jamming in different cities, any sort of software that has extras like that, would be fun... it could be a means to prolong your jam if it's getting boring or something. You could try different sounds or just mess around with it. But there's a practicality to the features too."*

Throughout all sessions, the musicians had also been providing feedback on improving the overall sound of the system, recommending preferred volume and reverb levels, and suggesting means to reduce any distortion. By Session 5, all of them were very pleased with how far the system had evolved, describing the sound as far "smoother" and more pleasant to the ear than it was at the start of our collaboration. For instance, when asked how they would gauge the changes in sound quality based on their previous suggestions, the guitarist explained:

*"It's definitely come a long way in terms of the quality of the sound that's coming through my ears. So that's the idea, I guess. It sounds good, so that's good."*

## 4.6 Additional Aggregated Results

**Analysis of Post-condition Questionnaires.** As noted earlier, musicians also completed post-condition questionnaires during each of the A/B/A tests. The questions were designed to assess a number of factors, such as the musicians' perceived sense of enjoyment, creativity and self-expression. Responses were tabulated and analyzed to determine the number of musicians for whom each of the system's features helped improve the factors listed above. As seen in Fig. 8, all features helped contribute to increased levels of enjoyment, with Musician Spatialization and Dynamic Reverb performing best in that regard. Furthermore, Track Panning contributed to an improvement in the musicians' sense of self-expression. Overall, however, creativity appeared to be the factor that least benefited from our system features, increasing only when Dynamic Volume was in use.



**Fig. 9.** Occurrences of positive and negative comments made under each major category in post-performance discussions.

**Analysis of Post-test Discussions.** All of our post-test discussions with the musicians were recorded and transcribed before a Qualitative Data Analysis (QDA) was performed. During a repeated coding process, comments were labelled and grouped, until three major categories emerged: Interaction, Sound Quality and Perceived Usefulness, and comments under each were tagged as being “positive” or “negative”.

As seen in Fig. 9, the number of positive comments for each category slowly improved throughout the sessions, with a particularly sharp increase in the “Interaction” and “Perceived Usefulness” categories seen during Session 5. We believe this to be, in large part, due to the nature of the session itself, as musicians were given the opportunity to try out the system features after all the feedback and suggestions they provided had been incorporated. In contrast, Fig. 9 also shows a steady decrease in the number of negative comments made for all three

categories. Together, these results indicate that we were successful in systematically incorporating the musicians' feedback into our system design. In the end, the band members found the system that evolved from our weekly sessions to be a vast improvement over its predecessor.

#### 4.7 Discussion

Our long-term deployment with the three-piece band was not only beneficial in allowing us to fine-tune EmbodiNet's existing features and introduce new ones, but it also helped us better understand the effects of embodied interaction on distributed performance. As Figs. 6 and 7 illustrate, the musicians made greater use of their local spaces when the Dynamic Volume and Dynamic Reverb features were in use. This is a marked improvement over traditional distributed performance systems, which, much like standard videoconferencing systems, do not encourage, or in some cases even support, a greater level of movement. Since Dynamic Volume and Dynamic Reverb also allow musicians to mutually affect one another's volume and reverb levels, their use also marks an increase in the sense of interplay among distributed musicians. In addition, Track Panning helped facilitate performance by helping the musicians maintain their rhythm with one another, while Dynamic Reverb added an aesthetic dimension to their sound while helping to reinforce the notion of a shared space. We also note that, with the exception of Musician Spatialization requiring a second explanation, the musicians experienced no difficulty in understanding and using our system features, and required little to no training. This helped further illustrate that, by designing them to capitalize on common gestures, embodied interactions can help enhance rather than detract from the higher-level task at hand.

While co-present performance, much like face-to-face communication, will always remain the gold standard, the musicians expressed that in a scenario where they were displaced, our system features would very much entice them to partake in an activity that they, like many other musicians, would otherwise likely not consider. Nonetheless, our long-term collaboration with the musicians also uncovered a number of shortcomings. First, some of EmbodiNet's features can only be experienced passively by seated musicians and, therefore, alternative controls must be designed to better accommodate such participants. In addition, with the exception of Dynamic Volume, the features did not necessarily help the musicians feel more creative. As we consider creative engagement to be integral to the musical experience, additional work is required in order to determine how EmbodiNet can better support such a quality. Finally, having established the practical aspects of seamless volume, reverb and mix adjustments, we would also like to explore the effects of providing musicians with controls of a more abstract or artistic nature.

Through EmbodiNet, we also hoped to examine how the shortcomings of distributed collaboration could be resolved in a manner that not only bridges the gap between the co-present and distributed contexts, but also serves to further enhance those aspects of the activity at hand that our users had deemed

most valuable. The underlying basis of telepresence research, the most prominent example of the “same time/different place” category of CSCW, has been to engender, as best as possible, a feeling of co-presence through the support of the non-verbal cues and gestures that are typically poorly conveyed between remote participants. The problem with such an approach is that it can only, at best, mimic co-location. Within the context of distributed performance, a parallel methodology is perhaps best illustrated through the breadth of research that aims to decrease latency and increase bandwidth as a means of facilitating musical collaboration. We argue, however, that the goal of distributed systems should not stop at simply mirroring their co-present counterparts. Distributed collaborative environments must, by nature, introduce a certain level of technology to offer their users support over even the most basic aspects of cooperation. As a result, we question whether participants stand to benefit from developers leveraging the technology at their disposal towards *augmenting* the activities these systems afford.

In our case, supporting audio sharing, the most elementary aspect of network performance, meant equipping each location with tools that musicians do not require under normal circumstances. With the addition of a Microsoft Kinect, we were able capitalize on the computing power necessary to make distributed performance a possibility, in a bid to present the network as a unique and appealing medium in its own right to the less technologically inclined musicians. In the end, our goal of augmenting an existing activity in a manner that utilizes existing, well-understood embodied interactions helped the musicians perceive distributed performance as a activity in which they would likely partake again, as expressed by members of the three-piece ensemble at the end of our long-term collaboration.

## 5 Conclusions

The on-going trend in the “same time/different place” category of Computer-Supported Cooperative Work has been to support, as best as possible, practices that might engender the feeling of co-presence (e.g., telepresence systems). In other words, the goal of such systems is to mimic co-present collaborative environments, in large part through support of the non-verbal cues and gestures that are often poorly conveyed between remote participants. We argue, however, that the goal of distributed systems should not simply stop at mirroring their co-present counterparts. Instead, such systems can leverage their underlying technology towards augmenting, and in turn, perhaps even improving, existing activities.

Through our experience with EmbodiNet, we tested this philosophy within the context of distributed performance, a domain we view as a unique extension of the “same time/different place” category of CSCW systems. EmbodiNet augments distributed performance by capitalizing on embodied interpersonal interactions among remote musicians. Through the use of five unique features, Dynamic Volume, Dynamic Reverb, Track Panning, Mix Control and Musician

Spatialization, musicians are able to seamlessly create and alter individualized mixes mid-performance, simply by moving around their space. Furthermore, by responding to changes in position and orientation, our performance environment allows musicians to utilize its features without having to detach themselves from the primary task of music-making.

The latest implementation of EmbodiNet was the result of a long-term collaboration and experiment with a three-piece band. Questionnaires collected throughout the collaboration proved that our system helped enhance the musicians' sense of enjoyment and self-expression. Position and orientation data indicated that the musicians took advantage of EmbodiNet's features, leading to an increased sense of interplay and spontaneity in spite of their remoteness. Furthermore, qualitative analysis of our discussions with the band members has shown that they found EmbodiNet to be practical, and that they would likely use it again in the future. By augmenting distributed performance, EmbodiNet, helped present such a domain as different from its co-present counterpart, yet appealing in its own right. As a result, we believe that the design of CSCW systems may benefit from a similar examination of how the collaborative activities they support could come to offer advantages that cannot be offered in traditional co-located contexts.

## References

1. Ackerman, M.S.: The intellectual challenge of CSCW: the gap between social requirements and technical feasibility. *Hum.-Comput. Interact.* **15**(2), 179–203 (2000)
2. Adalgeirsson, S.O., Breazeal, C.: MeBot: a robotic platform for socially embodied presence. In: Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction, pp. 15–22 (2010)
3. Álvaro, B., Kaltenbrunner, M.: Public sound objects: a shared musical space on the web. In: Proceedings of the International Conference on Web Delivering of Music, pp. 9–15. IEEE Computer Society Press (2002)
4. Antle, A.N., Corness, G., Bevans, A.: Springboard: designing image schema based embodied interaction for an abstract domain. In: England, D. (ed.) Whole Body Interaction. Human-Computer Interaction Series, pp. 7–18. Springer, London (2011)
5. Antle, A.N., Corness, G., Droumova, M.: Human-Computer-Intuition? Exploring the cognitive basis for intuition in embodied interaction. *Int. J. Arts Technol.* **2**(3), 235–254 (2009)
6. Bakker, S., Antle, A.N., Van Den Hoven, E.: Embodied metaphors in tangible interaction design. *Pers. Ubiquit. Comput.* **16**(4), 433–449 (2012)
7. Bannon, L.J., Schmidt, K.: CSCW: four characters in search of a context. In: Bowers, J.M., Benford, S.D. (eds.) Studies in Computer Supported Cooperative Work, pp. 3–16. North-Holland Publishing Co., Amsterdam (1991)
8. Braasch, J., Valente, D.L., Peters, N.: Sharing acoustic spaces over telepresence using virtual microphone control. In: Audio Engineering Society 123rd Convention (2007)

9. Carôt, A., Rebelo, P., Renaud, A.B.: Networked music performance: state of the art. In: Audio Engineering Society Conference 30th International Conference: Intelligent Audio Environments (2007)
10. Carôt, A., Werner Fischinger, C.: Towards a Comprehensive Cognitive Analysis of Delay-Influenced Rhythmic Interaction, pp. 473–476. MPublishing, University of Michigan Library, USA (2009)
11. Carstensen, P.H., Schmidt, K.: Computer supported cooperative work: new challenges to systems design. In: Itoh, K. (ed.) *Handbook of Human Factors*, pp. 619–636 (1999)
12. Carstensen, P.H., Schmidt, K.: Computer Supported Cooperative Work: New Challenges to Systems Design (1999)
13. Chafe, C., Wilson, S., Leistikow, A., Chisholm, D., Scavone, G.: A simplified approach to high quality music and sound over IP. In: Proceedings of the COST G-6 Conference on Digital Audio Effects, pp. 159–164 (2000)
14. Chew, E., Sawchuk, A., Tanoue, C., Zimmermann, R.: Segmental tempo analysis of performances in performer-centered experiments in the distributed immersive performance project. In: Proceedings of International Conference on Sound and Music Computing (2005)
15. Corness, G., Schiphorst, T.: Performing with a system's intention: embodied cues in performer-system interaction. In: Proceedings of the 9th ACM Conference on Creativity and Cognition, pp. 156–164 (2013)
16. Dourish, P., Bly, S.: Portholes: supporting awareness in a distributed work group. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 541–547 (1992)
17. Eisert, P.: Immersive 3-D video conferencing: challenges, concepts, and implementations. In: Proceedings of the International Society for Optical Engineering, pp. 69–79 (2003)
18. El-Shimy, D., Hermann, T., Cooperstock, J.R.: A reactive environment for dynamic volume control. In: Proceedings of the International Conference on New Interfaces for Musical Expression (2012)
19. El-Shimy, D.: Exploring user-driven techniques for the design of new musical interfaces through the responsive environment for distributed performance. Ph.D. thesis, McGill University (2014). <http://www.cim.mcgill.ca/~dalia/papers/el-shimy-thesis.pdf>
20. El-Shimy, D., Cooperstock, J.R.: Reactive environment for network music performance. In: Proceedings of the International Conference on New Interfaces for Musical Expression (2013)
21. Fencott, R., Bryan-Kinns, N.: Computer musicking: HCI, CSCW and collaborative digital musical interaction. In: Holland, S., Wilkie, K., Mulholland, P., Seago, A. (eds.) *Music and Human-Computer Interaction*. Springer Series on Cultural Computing, pp. 189–205. Springer, London (2013)
22. Grudin, J.: Why CSCW applications fail: problems in the design and evaluation of organizational interfaces. In: Proceedings of the 1988 ACM Conference on Computer-Supported Cooperative Work, pp. 85–93 (1988)
23. Ishii, H.: TeamWorkStation: towards a seamless shared workspace. In: Proceedings of the 1990 ACM Conference on Computer-supported Cooperative Work, pp. 13–26 (1990)
24. Ishii, H., Kobayashi, M., Grudin, J.: Integration of interpersonal space and shared workspace: clearboard design and experiments. *ACM Trans. Inf. Syst.* **11**(4), 349–375 (1993)

25. Johansen, R.: GroupWare: Computer Support for Business Teams. The Free Press, New York (1988)
26. Kapur, A., Wang, G., Davidson, P., Cook, P.R.: Interactive network performance: a dream worth dreaming? *Organised Sound* **10**, 209–219 (2005)
27. Lee, M.K., Takayama, L.: “Now, I Have a Body”: uses and social norms for mobile remote presence in the workplace. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 33–42 (2011)
28. Levisohn, A., Schiphorst, T.: Embodied engagement: supporting movement awareness in ubiquitous computing systems. *Ubiquit. Learn.: Int. J.* **3**(4), 97–112 (2011)
29. Loke, L., Robertson, T.: Design representations of moving bodies for interactive, motion-sensing spaces. *Int. J. Hum.-Comput. Stud.* **67**(4), 394–410 (2009)
30. Ravasio, P., Guttormsen-Schar, S., Tscharte, V.: The qualitative experiment in HCI: definition, occurrences, value and use. *Trans. Comput.-Hum. Interact.* 1–24 (2004)
31. Rebelo, P., Renaud, A.B.: The frequencyliator: distributing structures for networked laptop improvisation. In: Proceedings of the International Conference on New Interfaces for Musical Expression, pp. 53–56 (2006)
32. Renaud, A.: Dynamic cues for network music interaction. In: 7th Sound Music and Computing Conference (2010)
33. Rodden, T., Blair, G.: CSCW and distributed systems: the problem of control. In: Bannon, L., Robinson, M., Schmidt, K. (eds.) Proceedings of the Second European Conference on Computer-Supported Cooperative Work, pp. 49–64. Springer, Netherlands (1991)
34. Schroeder, F., Renaud, A.B., Rebelo, P., Gualda, F.: Addressing the network: performative strategies for playing apart. In: International Computer Music Conference (2007)
35. Schroeder, F., Pedro, R.: Sounding the network: the body as disturbant. *Leonardo Electronic Almanac* **16**(4–5), 1–10 (2009)
36. Schuett, N.: Interconnected musical networks: bringing expression and thoughtfulness to collaborative music making. Ph.D. thesis, Stanford Center for Computer Research in Music and Acoustics (2002). [https://ccrma.stanford.edu/groups/soundwire/publications/papers/schuett\\_honorThesis2002.pdf](https://ccrma.stanford.edu/groups/soundwire/publications/papers/schuett_honorThesis2002.pdf)
37. Sellen, A., Buxton, B., Arnott, J.: Using spatial cues to improve videoconferencing. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 651–652 (1992)
38. Sirkin, D., Ju, W.: Consistency in physical and on-screen action improves perceptions of telepresence robots. In: Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction, HRI 2012, pp. 57–64 (2012)
39. Tanaka, A., Bongers, B.: Global string: a musical instrument for hybrid space. In: Proceedings of Cast 2001: Living in Mixed Realities (2001)

# Preference Elicitation and Negotiation in a Group Recommender System

Jesús Omar Álvarez Márquez<sup>(✉)</sup> and Jürgen Ziegler

University of Duisburg-Essen, Duisburg, Germany  
[{jesus.alvarez-marquez, juergen.ziegler}@uni-due.de](mailto:{jesus.alvarez-marquez, juergen.ziegler}@uni-due.de)

**Abstract.** We present a novel approach to group recommender systems that better takes into account the social interaction in a group when formulating, discussing and negotiating the features of the item to be jointly selected. Our approach provides discussion support in a collaborative preference elicitation and negotiation process. Individual preferences are continuously aggregated and immediate feedback of the resulting recommendations is provided. We also support the last stage in the decision process when users collectively select the final item from the recommendation set. The prototype hotel recommender Hootle is developed following these concepts and tested in a user study. The results indicate a higher overall satisfaction with the system as well as a higher perceived recommendation quality when compared against a system version where no negotiation was possible. However, they also indicate that the negotiation-based approach may be more suitable for smaller groups, an aspect that will require further research.

**Keywords:** Group recommender system · Group preference elicitation · Negotiation · Decision making

## 1 Introduction

Over the recent years, recommender systems have proven beneficial in supporting users when selecting or buying items from large sets of alternatives [30]. Buying something in a virtual shop, deciding which film to watch or planning where to go on holidays can easily become a tedious task when solely relying on manual search and filtering techniques, which may lead to information overload and choice difficulties. Therefore, the importance of recommender systems has increased fast in the last years, being now used widely throughout the internet. While the field of recommendations for single users has already been deeply explored, the same cannot be said about group recommender systems. Even though a significant number of group recommenders have been developed in the past years [5, 18], there is still a range of issues which have not been sufficiently investigated so far.

Most group recommending approaches rely on existing user profiles which are either aggregated into a single group profile (model aggregation) before generating group recommendations, or which are used for calculating individual recommendations that are subsequently aggregated, using a variety of different strategies (recommendation aggregation). However, while sufficient profile information is often not available

in the case of single users – either due to a cold start condition, or because users do not want their profile to be stored – this problem is even more pertinent for groups where the likelihood of each user having a stored profile that can be exploited by the recommender is relatively low. This is especially the case for ad hoc groups who gather spontaneously or who come from different organizational contexts. A further issue is the situational variability of the group members’ preferences. This is also a problem in single-user recommending, but is aggravated by the fact that the inherent heterogeneity of preferences in a group may be amplified due to different responses to the situational context. These issues ask for methods that can elicit group preferences on the fly and that can aggregate individual preferences in a manner that best suits the individual users as well as the group as a whole.

Solving the complex trade-off between the degree of satisfaction of individual users and the group as a whole is typically attempted by applying one out of set of fixed strategies, such as averaging the satisfaction of all group members or minimizing discomfort for the least satisfied user. However, fixed strategies do not take the dynamics of group settings and situational needs into account. In particular, the aspect of social interaction when moving towards a joint decision is typically not sufficiently supported in existing group recommenders.

In this paper, we propose a novel method that tries to approach group recommendations from the point of intersection of traditional group recommenders and group decision making theory, allowing users to collaboratively create a preference model (thus addressing collaborative preference elicitation [28]), from which recommendations are generated. In this process, group interaction can happen at two (tightly intertwined) stages: (1) users can online discuss and negotiate preferences stated by others, and (2) they can discuss and rate items taken from the recommendation set to arrive at a final consensus decision.

Following the idea that computer-mediated discussion groups have more equal member participation [32], the goal is to avoid unfair situations in which some users might not be satisfied with the items proposed by the system. Our system supports remote online negotiation, although the approach can also be adapted to co-located settings. Each user can specify an individual preference model by freely adding desired features, using an explicit preference elicitation approach [27]. The individual preferences are then aggregated to form the group preference model and to determine an initial set of recommendations. All members’ preferences, as well as the group aggregation, are visible to the participants. Most importantly, individual preferences can then be negotiated in a system-supported manner: by group discussion, members may thus be able to convince other users to modify their preferences, so the group model changes to better match all members’ desires. Recommendations are continuously calculated and updated when the group preferences change, thus allowing users to immediately see the effect of their actions. Different mechanisms are provided for discussing and reaching an agreement, both for the creation of a group preference model and for the final item selection.

In the following, we first survey related research before presenting the conceptual aspects of our approach. We then describe the prototype implementation *Hootle* and its user interface design. We report on a user study we performed with groups of different sizes and conclude by summarizing our work and outlining future work.

## 2 Related Work

While the field of recommending items for single users has already received a great deal of attention in recent research, leading to quite effective recommendation methods, recommender systems for groups are, in comparison, a still less deeply investigated area. Various group recommender systems have been developed over the recent years, starting from early systems such as *MusicFX* [19], a group music recommender, that use different approaches for generating recommendations [5, 12]. However, there are still many open research questions concerning, for example, the best approach to aggregating individual preferences, techniques for responding to the situational needs of the group, or supporting the social interaction processes in the group for converging on a joint decision.

To structure the wide range of different aspects involved in group recommending, [14] suggest a design space comprising the dimensions preference input (including dynamic aspects), process characteristics, group characteristics, and (presentation of) output. In the process dimension, an important aspect is how individual, possibly conflicting preferences can be merged to obtain recommendations that best fit the group as a whole. Although different approaches in group recommenders gather and represent users' preferences in different ways, they commonly use one of two schemas [12]:

**Aggregation of Item Predictions for Individual Users (Prediction Aggregation).** This approach assumes that for each item, it is possible to calculate a user's satisfaction, given the user's profile. Then, using the calculated predictions and making use of some specific aggregation strategy, items are sorted by the group's overall satisfaction. In [9] a video recommender that uses this strategy is described; also *PolyLens* [26], a system that suggests movies to small groups of people with similar interests, based on the personal five-star scale ratings from *Movielens* [8] uses this method.

**Construction of Group Preference Models (Model Aggregation).** Instead of predicting matching items for each user, the system uses information about individual members to create a preference model for the group as a whole. Recommendations are generated by determining those items that best match the group model. The number of possible methods for creating the group's model is even bigger than it is for prediction aggregation strategies. For example, in *Let's Browse* [15] the group preference model can be seen as an aggregation of individual preference models. In *Intrigue* [1, 2] (which recommends sightseeing destinations for heterogeneous groups of tourists) the group preference model is constructed by aggregating preference models of homogeneous subgroups within the main group. *MusicFX* [19] chooses background music in a fitness center to accommodate members' preferences, also by merging their individual models. AGReMo [4] recommends movies to watch in cinemas close to a location for ad hoc groups of users, creating the group's preference model not only by individual model aggregation but also taking into account some specific group variables (e.g. time, weight of each member's vote). Furthermore, the *Travel Decision Forum* [10, 11] creates a group preference model that can be discussed and modified by the members themselves, aiming to non-collocated groups who are not able to meet face to face, allowing asynchronous communication.

Regardless of whether the aggregation is made before or after generating recommendations, an aggregation method that is appropriate for the specific group characteristics needs to be chosen. There are a number of voting strategies, empirically evaluated in [18], that have been used in actual group recommender systems. Some typical strategies (and systems using it) are:

- **Average strategy**, where the group score for an item is the average rating over all individuals (*Intrigue*, *Travel Decision Forum*).
- **Least misery strategy**, which scores items depending on the minimal rating it has among group members (*Polylens*, *AGReMo*).
- **Average without misery strategy**, consisting in rating items using an average function, but discarding those where the user score is under a threshold (*MusicFX*, *CATS* [20–23]).
- **Median strategy**, which uses the middle value of the group members' ratings (*Travel Decision Forum*).

On another dimension, the question of preference elicitation has to be solved, which is concerned with how the user-specific preference information needed to generate recommendations is obtained. One approach is to let users rate a number of items in advance and to derive preferences from this set of ratings. *AGReMo*, for instance, requires group members to create their own model of individual preferences before the group meeting takes place by rating movies that they already saw. In *Travel Decision Forum* each participant starts with an empty preference form that has to be filled with the desired options, so group members define new preferences for each session. A more interactive approach, although for single user systems, is described in [17] which requires users to repeatedly choose between sets of sample items that are selected based on latent factors of a rating matrix. The techniques mentioned also address the cold-start problem when no user profile is available up-front but initially require some effort on the part of the user to develop a sufficiently detailed profile.

However, most preference elicitation techniques do not take group interaction into account. As pointed out in [16], to obtain adequate group recommendations it is not only necessary to model users' individual preferences, but also to understand how a decision among group members is reached. While research on group decision making [31] is concerned with collaboratively making choices, focusing on the social process and the outcome, these aspects have mostly not been addressed in the development of group recommender systems. The process of group decision making involves a variety of aspects, such as the discussion and evaluation of others' ideas, conflict resolution, and evaluating the different options that have been elaborated. Also interesting for our research is the concept of consensus decision-making [7], which seeks for an acceptable resolution for the whole group. Within this context, Group Decision Support Systems (GDSS) have emerged, that aim at supporting the various aspects of decision making [24, 25]. Only few recommender systems attempt to include aspects of group decision theory, for instance, by introducing automated negotiation agents that simulate discussions between members to generate group recommendations [3]. However, supporting the entire preference elicitation and negotiation process that may occur when users take recommender-supported decisions is, to our knowledge, not realized by current group recommenders.

Lastly, taking into account the social factor that is involved in group recommendation, one needs to contemplate the question whether a user would be willing to change personal preferences in favor of the group's desires, bringing up the importance of group negotiation. Again in the *Travel Decision Forum*, users are able to explore other members' preferences, with the possibility to copy them or propose modifications. The *Collaborative Advisory Travel System (CATS)* focuses on collocated groups of persons gathered around a multi-touch table. Recommendations are made by collecting critiques (users' feedbacks respecting recommended destinations) that can be discussed face to face, since the system gives visual support to enhance awareness of each other's preferences. The main difference between CATS and the system proposed here is that the former is focused in critiquing items once they have been recommended, while the latter allows negotiation already in the preference elicitation stage.

### 3 Preference Elicitation and Negotiation Method

The method developed involves an iterative process of specifying, discussing and negotiating preferences in a remote collaboration setting. Instead of only discussing recommendations produced based on user profiles, interaction among group members is supported right from the beginning of the preference elicitation process. The overall process comprises the following stages which are not meant as sequential steps but which can basically be performed in any order (algorithmic and interface details are described in the next chapter):

1. Users begin by selecting desired features from a set of attributes describing the items available. Since the feature sets may be very large (e.g. cities in our example hotel recommender, users can first search for the features they want and place them in a private area).
2. By moving a feature to the user's individual preference list, the feature becomes active and is visible to other group members. Several features can be placed and rank ordered according to the relevance they have for the user.
3. The individual feature lists are constantly aggregated in a common, ranked group preference list and the recommendations that best match the current group model are immediately generated and shown to the group.
4. Users can discuss preferences stated by others and negotiate them by using a 'petition' function, potentially trading in own preferences for features other users want. Based on the discussions and negotiations, users may change their preferences which is again immediately reflected in the group model and the resulting recommendations.
5. From the recommendations users can at any time select the item(s) they really like and propose them to the other participants who can accept them or propose alternatives. Also in this stage of the process, discussions are supported by the system.

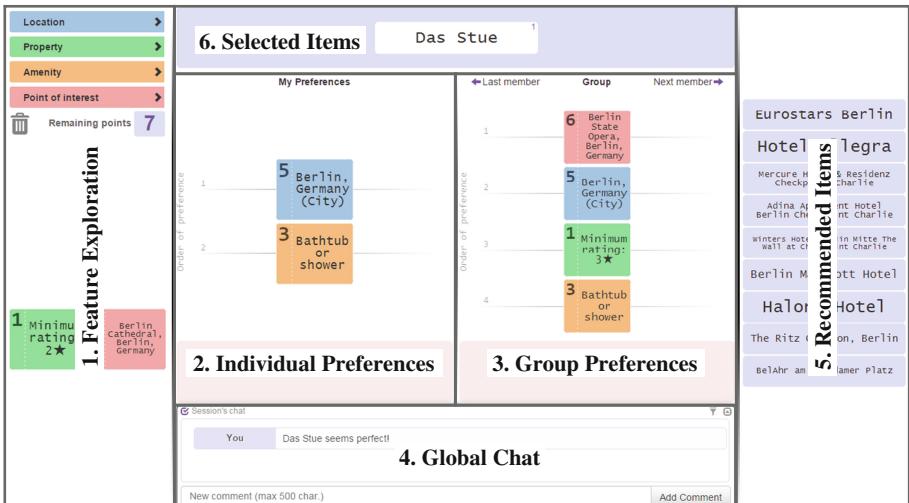
The closed loop interaction with immediate feedback in the group model and the recommendations increases participants' awareness of others' preferences and the effects their own preference changes have on the group results. The approach also entails aspects of critique-based recommenders since users can criticize or accept

proposed features or recommended items. In contrast to fully automated recommender system, users have a higher level of control over the process and can easily adapt it to their current situational needs and context.

## 4 Description of the System

To demonstrate our approach we designed and implemented a prototype group recommender system that employs content-based techniques. The system is in principle applicable in a wide range of application areas, such as candidate selection, requirements specification, or leisure activities, as long as it is possible to obtain the properties of the items to be recommended. For demonstration purposes, we chose hotel selection for group travel as application area and use an Expedia dataset consisting of 151.000 hotel entries with descriptive information.

Figure 1 shows a screenshot of the user interface, described as following:



**Fig. 1.** Areas of the interface.

- 1. Feature exploration.** This area consists of a set of defined filters that let users search for specific attributes and a space to store the selected ones. For example, filters could be location, facilities or nearby points of interest.
- 2. Individual preferences.** Features selected in area 1 can be added here by drag-and-drop, meaning that the user wants these features to be present (or excluded) in the recommended items (more details in the section about Individual Preferences). Users can also rank their preferences to express different levels of importance.

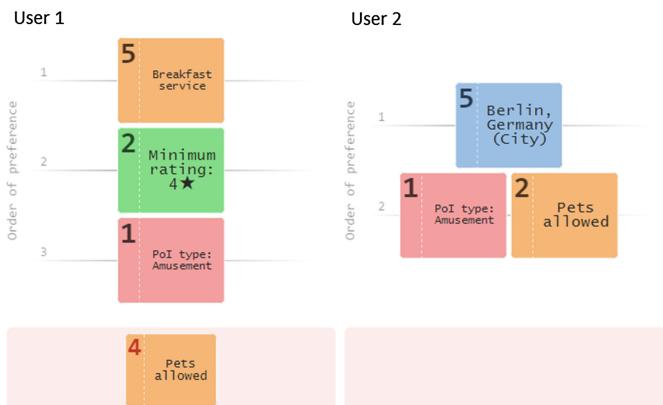
3. **Group preferences.** A ranked aggregation of all individual preferences is displayed in this area. It is also possible for users to navigate through the preferences of other participants here.
4. **Global chat.** In this section, the group can discuss arbitrary questions that come up in the decision process. Requests for preference changes (“petitions”) and comments about specific features can also be displayed here.
5. **Recommended items.** Here, the items that best match the current group preferences and their relative weight are shown. The list is constantly updated in real-time when users add or change features.
6. **Recommendations selected by users.** From the recommendations area, users can pick the items they like most, and place them here. This space works as a shared area, so each item added here is visible to all participants.

#### 4.1 Feature-Based Preference Elicitation

**Individual preferences** are defined by each group member by selecting features from the exploration area, where they can use different filters to locate them. Later, features can be placed into the user’s individual preference space. The system allows to specify both positive and negative features.

*Positive features.* A positive property means that a user wants it to be found in the recommendations. Users can specify an order of preference among positive attributes by dragging them to a higher or lower position in the list, which denotes the degree of importance that the user gives to each feature. Multiple features may have the same preference level.

*Negative features.* Negative properties are those that the user does not want to get as feature of the recommended items. They are placed inside a subspace within the



**Fig. 2.** Example of preference areas belonging to two different users. The ordered list represents the positive (desired) attributes, while the area at the bottom contains the negative (vetoed) ones. The cost of each attribute can be found at the top-left corner.

individual area (Fig. 2), called the veto area. Vetoed attributes have no preference order.

*Cost of features.* When users specify a large number of features as preferences, several problems may arise: first, it may be difficult to create meaningful integrated group preferences because the probability that features contradict each other increases, requiring more complex and longer negotiation processes. Second, users may over-specify their preferences making it difficult or impossible to calculate well-matching recommendations. We therefore decided to devise a mechanism that gently pushes users towards only specifying the features they really want.

For this purpose, a method for measuring the cost of each feature has been implemented. Each attribute has a related cost depending on how restrictive it is (i.e. how many items are left after using it as filter over the database). When a user selects a feature he or she pays for it from a limited budget. Users only have a number of tokens to exchange for attributes so they have to choose which ones are most important. This way, users selecting very restrictive features will only be able to create a small list of preferences as they will cost more tokens. It is also necessary to remark that the cost for positive attributes differs from the one for negatives. Positive attributes are more expensive the more restrictive they are; for negative features, more restrictiveness means less cost.

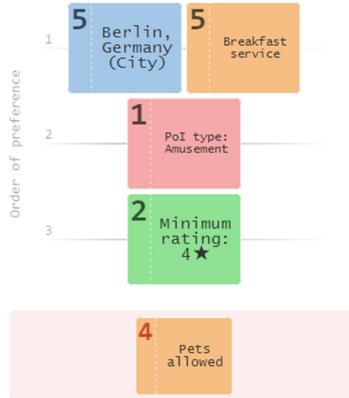
**Group Preferences.** While creating their individual preference lists, users can immediately see the overall results for the group. Inside the group preference area, an aggregation of all individual user preferences is displayed. This list is called the group preference list. The aggregation of individual preferences is performed using a variant of the Borda Count method, combined with rules regarding the vetoed attributes.

Borda Count is a voting method in which voters rank options or candidates in order of preference. In standard Borda Count, each option receives a score depending on its rank, and to obtain the aggregated score the points that each voter has given to it are summed up. In the case at hand, not only the rank of each option has been taken into account, but also its cost. When a user chooses to place a relatively expensive (restrictive) feature in the individual preference list, it is fair to think that the user cares more about this specific attribute. The equation used to calculate the aggregated score of an attribute  $i$  is presented in (1), where  $u$  is the number of group's members,  $n$  is the total number of different attributes used,  $p_{ij}$  is the preference value given to the attribute  $i$  by the user  $j$ ,  $c_i$  is the cost of the attribute  $i$  and  $\lambda$  is used to correct the importance of the cost (with  $\lambda = 0$  the result would be a standard Borda Count voting aggregation).

$$PAtt_i = \sum_{j=0..u} \left( \frac{1}{n} (n - p_{ij}) \right) + \frac{\lambda c_i}{n} \quad (1)$$

Attributes only receive points if users include them in their preferences. Finally, the group preference list is created by calculating the total score for each item and sorting them as usual (Fig. 3).

Vetoing a feature is a strong statement, it means that the person who stated it really does not want items with this feature. It would be desirable to avoid this feature, even if



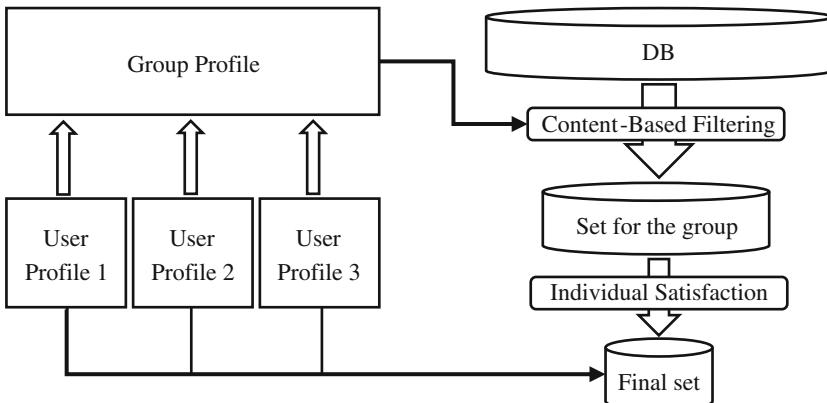
**Fig. 3.** Resulting preference setting for the group, using the individual lists shown at Fig. 2.

someone else in the group still wants it. Thus, vetoed attributes are removed from the group preference list and will not appear in any of the recommendations.

#### 4.2 Generating Recommendations

Based on the aggregated user preferences the system applies a content-based filtering method to generate recommendations (Fig. 4). In content-based filtering, items are described by a set of attributes, and each user has a profile of preferences indicating the item properties the user likes. In our case, the individual preference set in a session represents the full user profile, thus, the system is applicable in cold-start situations where no user profile exists yet.

To generate recommendations, group preferences are compared to the items' properties in order to find the best matching ones. First, the system removes all the items



**Fig. 4.** Scheme of the filtering process.

that contain a vetoed attribute. The remaining items receive a score based on how many positive features they match, their total score being the sum of their attributes' values. The value of each attribute comes defined by the Borda Count method previously described, so attributes with higher preference levels will give higher score values to the items containing them. For distance attributes (coordinates, regions or points of interest), the value they were assigned by the Borda Count is modified depending on how far an item is from the given feature (closer items obtain higher scores).

If the system would simply present the ten top scored items, it could happen that for some users whose attributes are lower in the group preference list, no good options are returned. Since the main purpose of the system is to provide a negotiation environment, it seems necessary to return a well-balanced set of items, in terms of member satisfaction. For this reason, a subset of items is extracted, within the already found, in a way that for each user there is at least one acceptable option, but giving at the same time importance to the items that satisfy the group as a whole. An item is considered acceptable for a participant when his/her satisfaction level concerning this option is higher than a given threshold. Satisfaction is calculated taking into account the individual preference model defined by a user, in a similar way an item's group score is calculated, but divided by the maximum points an item could receive (that is, when an item contains all the features a user wants). Finally, the selected items are presented to participants in the recommendation area of the screen (5 - Recommended Items in Fig. 1).

As said before, the system is applicable without requiring the prior availability of stored user profiles which is particularly beneficial in group contexts for the reasons mentioned earlier. However, in principle more complex and longer-term user profiles could be built if past choices were saved for future sessions. If this option was used and is acceptable for users, the interaction effort needed for specifying the desired features could be reduced, just specifying changes in the existing profile, and possibly increasing the precision of the recommendations.

### 4.3 Negotiation

User preferences are typically not a static phenomenon but are influenced by the situational context of the group and the social interaction that takes place within it. Users may also differ in the extent to which they have already formed their objectives at the beginning of the group process. They may react to preferences expressed by others, either accepting or rejecting them. They may also be willing to dispense with a desired feature if someone else in the group accepts one of their other preferences, thus embarking on a negotiation process with other group members. For these reasons, our system provides several functions that specifically support discussion, negotiation and consensus finding among group members.

**Communication.** Users need the possibility to express their opinions about the decision process as a whole as well as about specific preferences stated by others. To support these types of communication, two methods are implemented in the system.

*Discussion threads and global chat.* Each feature has its own discussion thread, which means that users can access it and say what they think about a specific property, keeping the comments organized by attribute. A global chat is also available, placed in area 4 displayed in Fig. 1. The global chat lets participants talk about arbitrary aspects of the current session, and also informs group members about recent updates in specific comment threads.

*Petitions.* Petitions are requests such as removing a feature or changing its rank. It is not possible to request the addition of an attribute, as adding a feature to one's individual list is already an implicit petition to the rest of the group: every user wants the others to adopt the same preferences as he/she has, since this would increase the fit of the recommendations with this user's wishes.

**Finding and Resolving Conflicts.** Conflicts appear when two or more participants want features that contradict each other. Several mechanism help to resolve such situations. First, users can explore the individual preferences of other participants and discuss them if a conflict occurs.

Second, once a set of recommendations is presented, users can access information about each item recommended. Also, those entries in the group preference list that are not fulfilled by an item are highlighted in that list. Thus, when a user likes a recommendation, he/she can see the preferences that are in conflict with it and try to change the opinion of the members who added them.

Finally, for each recommendation, the calculated grade of satisfaction of each user can be displayed in a spider diagram, so the group may choose items that are more balanced with respect to the members' individual desires (i.e. are less conflictive).

**Proposing Items.** From the recommendation area, users are able to express their approval for a specific recommended item by placing it into the “recommendations selected by users” space (area 6 in Fig. 1). This step shows the group that one user likes a recommendation and proposes it as option. The other participants now can accept it as a good option, reject it or just ignore it, waiting for more proposals to show up.

#### 4.4 Repeat and Decide

The “*adding features-get recommendations-negotiate*” cycle can be repeated several times, narrowing down the recommendations given with each new iteration, until the group reaches agreement. If and when consensus is reached, however, is something that only the group itself is able to decide. As has been said in the previous section, users can add items that they like into a shared area, so the others can express their acceptance about it. For some groups, the item to be finally selected may be the one that is accepted by more than fifty percent of the members; in other cases, there may be situations where all users have accepted an item except one who finds it unsatisfactory. While a fixed group recommendation strategy, for example, a ‘least misery’ approach that might seem applicable in the latter case, would always try to satisfy user needs in one prescribed manner, we believe that the system cannot generally resolve such decision problems. Although the system provides tools for preference specification,

discussion and acceptance measuring, it is up to the users to decide whether a recommendation fits their needs or not and to make the final choice.

## 5 Evaluation

To evaluate our approach, we performed a user study with several groups comprising between three and five users. We did not consider larger groups at this point because we believe this group size to be typical for the application domain chosen which is selecting a hotel for a joint leisure or business trip. Also, Hootle, our Web-based prototype implementation of the approach, while still work in progress, is stable enough to support this group size but still has to be tested for larger-scale trials. The main objectives of this study were to determine the usability of the approach and the quality of the resulting recommendations, as well as, more specifically, to analyze the impact of the cooperative preference elicitation and negotiation tools developed.

### 5.1 Setting and Experimental Tasks

To assess whether the preference elicitation, negotiation and recommendation methods developed benefit group decision processes, we tested two different versions of the system where one served as baseline for comparison. While one system version provided the full set of functions described including group discussion support (hereafter version D – Discussion), we restricted the second version to specifying preferences and calculating recommendations (version ND – No Discussion), similar to a conventional group recommender system, but still offering the possibility to specify preferences in an ad hoc manner without using existing user profiles. We decided against using an existing alternative group recommender for comparison because the systems would have differed in too many aspects, making it difficult to pinpoint the specific benefits of the proposed innovations. In both cases, we make use of a hotel database provided by Expedia with 151,000 entries. For each hotel, a full description and a set of attributes, including property and room amenities (within a total of 360 possibilities), locations (258,426) and points of interest nearby (94,512) was available. We deliberately decided to focus the negotiation and decision process on the objective properties of the items, excluding price information which would have opened up additional questions concerning economic concerns and behavior in the test groups. This aspect, however, will be subject of future research.

We prepared two types of task scenarios with different levels of complexity:

- In an ‘introductory’ task, the group was instructed to select a hotel knowing beforehand some common, desired attributes, as well as the location of the hotel. This task also served as a training session for the application, to allow participants to explore the functions and possibilities the system supplies. Two scenarios for this task were presented:
  - Your group will be participating at a conference in Berlin. As the conference always provides lunch and dinner, you just need to find a hotel including breakfast. Your conference takes place near the Brandenburg Gate.

- Your group wants to enjoy some days on the beach. You already decided to go to an apartment, as you want to prepare meals on your own. Everyone loves Spain so you also decided to go to Marbella.
- In the ‘open’ task which was always performed after the introductory task, only unspecific instructions were given to the group such as “Find a place to stay during summer vacation”. The possible scenarios were:
  - It is summertime. You and your friends really need to get out of the daily routine. Discuss where to stay.
  - Your group wants to do some kind of city trip. Where are you going to?

To avoid the problem that in a test situation, participants do not bring with them the objectives and preferences they would have in a real-life decision situation, or might comply too quickly with the wishes of other participants, we tried to artificially induce different backgrounds and objectives for each group member. For this purpose, we created a set of role cards for the second task, depending on the scenario used. With this method, we expected to generate conflicts and discussion when randomly distributing the role cards among group’s members. As an example, the role cards for the first scenario in task 2 were (abbreviated here):

1. You’re a sport addict. You like to eat healthy and don’t trust in hotel food. You hate giant hotels and prefer small pensions or camping sites.
2. You’re allergic to nearly everything. Vacation at a camping site would be like a death sentence to you. You prefer the pool over the sea. You don’t want to do anything so you prefer all inclusive.
3. You like to go for long hikes. You’re fascinated by mountains. You don’t want to cook but you won’t be there during the day so you just need breakfast and dinner.
4. You’re into cultural things. If you go on vacation, you want to see things. You also like to go out for dinner so breakfast only would totally fit your needs.
5. You like to party. As you won’t be able to prepare your own food, there should be someone who helps you with this. More important is the location of your hotel. Nobody wants to walk for an eternity to go clubbing.

## 5.2 Method

A total of 48 students were recruited as participants (5 male, 43 female, average age of 20.94,  $\sigma$  5.018), distributed in groups of different sizes: 4 groups of 3 persons (12), 4 groups of 4 persons (16) and 4 groups of 5 persons (20). Two groups of each size ran a full version of the system (D), while the other two groups tested the version without negotiation support (ND). Since the system is Web-based, all users were provided with a normal desktop computer with a display screen of 21 in and running the same browser. They sat in a large lab room but were separated from each other and instructed to only communicate via the means provided by the system.

Each group first received a brief introduction to the system which was dependent on whether the negotiation support was turned on or off for the group. After a brief trial, they were asked to work on the two decision tasks, always in the order introductory

task – open task. Before beginning the second task, they all received randomly one of the role cards.

For the groups using version D, a task was considered complete when they reached consensus about their preferred hotel or when they decided that it was not possible to find agreement. Since the groups with version ND were not able to communicate at all, their job consisted in defining their own preference model and, when the whole group had done this, each user separately selected a hotel from the resulting set of recommendations.

The first task including the explanation of the system was limited to a maximum of 40 min. As the explanation was no longer necessary, the second task, although more complex, should also be completed during this time.

After completing both tasks, participants were asked to fill in a questionnaire regarding aspects such as the quality of the recommendations or the ease-of-use of the system, using a 1-5 scale. The questionnaire comprised the SUS items [6] to compare the system against a well-established baseline as well as items from two recommender-specific assessment instruments (User experience of recommender systems [13] and *ResQue* [29]). The recommender-specific items were measuring mainly the constructs *user-perceived recommendation quality*, *perceived system effectiveness*, *interface adequacy*, and *ease of use*.

### **5.3 Results and Discussion**

All tasks were finished within the allotted time. The D and ND groups differ on a considerable number of criteria. The members in ND groups were not able to choose the same hotel in a single instance. In two of these cases, some users couldn't even find a hotel that they liked when realizing the open task. On the other hand, all groups with version D were able to choose one unique hotel in both tasks, despite starting the

**Table 1.** Results of the questionnaire (all the D/ND differences  $p > 0.05$ , effects of group size were significant).

System version		No discussion				Discussion			
Group Size		3	4	5	Avg.	3	4	5	Avg.
Overall satisfaction	m	3.40	3.00	3.70	3.39	4.33	4.00	3.60	3.92
Would recommend it	$\sigma$	0.54	1.20	0.48	0.83	0.51	0.53	0.96	0.77
Would use it again	m	3.20	2.38	3.30	2.96	3.50	3.25	3.30	3.33
	$\sigma$	1.30	1.06	0.67	1.02	0.83	0.70	1.06	0.86
Would use it again	m	2.40	2.50	3.10	2.74	3.17	3.13	3.00	3.08
	$\sigma$	0.89	0.92	1.10	1.01	0.75	0.99	0.66	0.77
Would use it frequently	m	1.60	1.88	2.30	2.00	2.67	2.75	2.70	2.71
	$\sigma$	0.54	0.64	0.67	0.67	0.81	1.04	0.94	0.90
Recommendations were well chosen	m	3.20	3.38	3.80	3.52	4.33	3.38	4.00	3.88
	$\sigma$	0.83	0.74	0.78	0.79	0.51	0.74	0.47	0.68

process with strongly different individual preferences. To achieve this joint decision, users had to iterate several times through the “*adding features-get recommendations-negotiate*” cycle, as well as to renounce some desired features due to the influence exerted by other members through discussions and petitions.

In terms of overall usability, both system versions received a SUS score which can be considered as borderline good with no differences between the two systems (ND = 68, D = 69). We performed a  $2 \times 3$  ANOVA with system version and group size as independent variables and questionnaire item scores as dependent variables. Most item responses did not show significant differences between the two system versions which may be due to the limited number of groups tested. In Table 1, we list some of the results that were significant at a .05 level. Users in the discussion condition were overall more satisfied with the system, are more likely to recommend it to others and would be willing to use the system again and also more frequently. Also, the accuracy of the recommendations was rated higher in the discussion groups. While these results speak in favour of the discussion version, there appears to be an interesting interaction effect between system versions and group size. Generally, satisfaction and willingness to use and recommend the system tend to be higher for the small groups than the large groups when discussion is available. Concerning recommendation quality, the largest group had the highest ratings in the no-discussion condition while this is reversed in the discussion condition where the smallest group had the highest rating. This picture is somewhat blurred by the fact that the medium-sized groups (4 persons) had the largest variability so there is no clear relation between group size and these variables.

For the remaining questionnaire items (which we cannot report here fully due to space limitations) there is a tendency in favour of the discussion version both in the items related to usability and acceptance of the system as well as concerning the fit of the recommendations and the ease with which a matching hotel could be found.

The time needed to come to a decision differed significantly between the introductory task and open task (13,500 vs. 26,333,  $p = 0.05$ ). Results concerning negotiation behavior are listed in Table 2: both individual changes and number of petitions increase with group size. In relation with Table 1, it may be concluded that users in small groups are generally more satisfied because they were able to select more preferences for themselves and made less changes in their individual lists (keeping their initial wishes).

*Discussion:* The results of this study can only give a first indication of how well the proposed approach works in comparison to other techniques and in different group

**Table 2.** Objective results (lower and upper bounds at 95 % confidence interval).

	3 Participants			4 Participants			5 Participants		
	m	LB	UB	m	LB	UB	m	LB	UB
Time	21	31	10,9	17,2	7,23	27,2	21,5	11,4	31,5
Pref. Sel./Part.	3	2,37	3,62	2,31	1,76	2,85	2,80	2,31	3,28
Ind. Changes/Part.	7,33	1,79	12,8	10,1	4,64	15,7	13,1	7,61	18,6
Petitions/Part.	0,66	0	1,65	0,68	0	1,54	1,95	1,18	2,71

contexts. We can see significant advantages for our approach of including discussion and negotiation features in a group recommender in some relevant items, as well as a tendency in favour of the system in the majority of other items. However, it appears that the system may be more useful in small groups. This may be due to several factors: first, as larger groups require more communication and negotiation to obtain an acceptable end results, this may increase the complexity of the task and the interaction effort. This may be true for other group decision making systems as well but will require further research. A second factor may be artificially created by the experimental method used. Since users were instructed to play the roles described in their respective role cards, the diversity of preferences increased with group size, possibly making it more difficult to make sense of the diverse standpoints and to lead the negotiation towards a joint group decision. This may not be the case in typical real world settings where group members' viewpoints may be more homogenous due to the prior history of the group. Also, the role card method can only be taken as an approximation of a real situation. In any case, the observed tendencies raise interesting general questions concerning test scenarios for evaluating group recommender systems.

## 6 Conclusions and Outlook

We have presented a novel approach to group recommending that provides more interactive control over the recommendation process than typical group recommenders and that does not require the prior availability of the group members' preference profiles, taking into consideration cold-start situations and potential privacy concerns. Most importantly, the method provides discussion and negotiation support in a collaborative preference elicitation and negotiation process. Individual preferences are aggregated in a group preference profile which is immediately updated when users change preferred features or their relevance level. Also, the resulting recommendations are continuously recalculated when group preferences change, and are always visible to the whole group. Since producing recommendations constitutes just an intermediate step in the group decision process, we also support group interaction in the final decision steps where the group needs to find consensus about the item finally selected.

The proposed technique provides much higher flexibility and responsiveness to situational needs than the fixed strategies typically used in group recommenders. While this research has focused on specifying preferences in an ad hoc fashion, the method can easily be extended by storing and re-using user profiles, thus reducing interaction effort to simply adapting an existing profile. Since the preferences of other users and resulting group preferences as well as the recommendations that match this profile are always visible, participants' awareness of individual and group views and of the effects of their preference settings is increased.

Based on these concepts, we developed the prototype hotel recommender Hootle and tested it in a user study. The results indicate a higher overall satisfaction with the system as well as a higher perceived recommendation quality when compared against a system version where no discussion was possible. However, we also saw an indication of an interaction effect between group size and the two system versions which suggests that the negotiation-based approach may be more suitable for smaller groups. Whether

this effect is due to the increased communication effort in larger groups, or may be dependent on the experimental scenarios used in the study is still an open question.

In future work, we aim at investigating the effects of group size more deeply and at optimizing the system to better scale for larger groups. A further work item is to consider alternative aggregation functions that may perform better than the Borda Count variant currently used. Finally, we aim at further improving the user experience with respect to the discussion and decision making features implemented. Also, more extensive empirical studies are planned, addressing also domains other than hotel selection.

## References

1. Ardissono, L., Goy, A., Petrone, G., Segnan, M.: A multi-agent infrastructure for developing personalized web-based systems. *ACM Trans. Internet Technol.* **5**(1), 47–69 (2005)
2. Ardissono, L., Goy, A., Petrone, G., Segnan, M., Torasso, P.: Intrigue: personalized recommendation of tourist attractions for desktop and hand held devices. *Appl. Artif. Intell.* **17**(8–9), 687–714 (2003)
3. Bekkerman, P., Kraus, S., Ricci, F.: Applying cooperative negotiation methodology to group recommendation problem. In: Proceedings of Workshop on Recommender Systems in 17th European Conference on Artificial Intelligence (ECAI 2006), pp. 72–75. Citeseer (2006)
4. Beckmann, C., Gross, T.: Towards a group recommender process model for ad-hoc groups and on-demand recommendations. In: Proceedings of the 16th ACM International Conference on Supporting Group Work, pp. 329–330. ACM (2010)
5. Boratto, L., Carta, S.: State-of-the-art in group recommendation and new approaches for automatic identification of groups. In: Soro, Alessandro, Vargiu, Eloisa, Armano, Giuliano, Paddeu, Gavino (eds.) *Information Retrieval and Mining in Distributed Environments*. SCI, vol. 324, pp. 1–20. Springer, Heidelberg (2010)
6. Brooke, J.: SUS-A quick and dirty usability scale. *Usability Eval. Ind.* **189**, 194 (1996)
7. Hartnett, T.: Consensus-Oriented Decision-Making: the CODM Model for Facilitating Groups to Widespread Agreement. New Society Publishers, Gabriola Island (2011)
8. Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.: Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst. (TOIS)* **22**(1), 5–53 (2004)
9. Hill, W., Stead, L., Rosenstein, M., Furnas, G.: Recommending and evaluating choices in a virtual community of use. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 194–201 (1995)
10. Jameson, A.: More than the sum of its members: challenges for group recommender systems. In: Proceedings of the Working Conference on Advanced Visual Interfaces, pp. 48–54. ACM (2004)
11. Jameson, A., Baldes, A., Kleinbauer, T.: Two methods for enhancing mutual awareness in a group recommender system. In: Proceedings of the Working Conference on Advanced Visual Interfaces, pp. 447–449. ACM (2004)
12. Jameson, A., Smyth, B.: Recommendation to groups. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) *Adaptive Web 2007. LNCS*, vol. 4321, pp. 596–627. Springer, Heidelberg (2007)
13. Knijnenburg, B.P., Willemse, M.C., Gantner, Z., Soncu, H., Newell, C.: Explaining the user experience of recommender systems. *User Model. User-Adap. Inter.* **22**(4–5), 441–504 (2012)
14. Kompan, M., Bielikova, M.: Group recommendations: survey and perspectives. *Comput. Inform.* **33**(2), 446–476 (2014)

15. Lieberman, H., Van Dyke, N., Vivacqua, A.: Let's browse: a collaborative browsing agent. *Knowl.-Based Syst.* **12**(8), 427–431 (1999)
16. Liu, X., Tian, Y., Ye, M., Lee, W.: Exploring personal impact for group recommendation. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, pp. 674–683. ACM (2012)
17. Loepf, B., Hussein, T., Ziegler, J.: Choice-based preference elicitation for collaborative filtering recommender systems. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2014), pp. 3085–3094. ACM, New York (2014)
18. Masthoff, J.: Group Modeling: Selecting a Sequence of Television Items to Suit a Group of Viewers, pp. 93–141. Personalized Digital Television. Springer, The Netherlands (2004)
19. McCarthy, J.F., Anagnost, T.D.: MusicFX: an arbiter of group preferences for computer supported collaborative workouts. In: Proceedings of the 1998 ACM Conference on Computer Supported Cooperative Work, pp. 363–372. ACM (1998)
20. McCarthy, K., McGinty, L., Smyth, B.: Case-based group recommendation: compromising for success. In: Weber, R.O., Richter, M.M. (eds.) ICCBR 2007. LNCS (LNAI), vol. 4626, pp. 299–313. Springer, Heidelberg (2007)
21. McCarthy, K., McGinty, L., Smyth, B., Salamó, M.: The needs of the many: a case-based group recommender system. In: Roth-Berghofer, T.R., Göker, M.H., Güvenir, H.A. (eds.) ECCBR 2006. LNCS (LNAI), vol. 4106, pp. 196–210. Springer, Heidelberg (2006)
22. McCarthy, K., McGinty, L., Smyth, B., Salamo, M.: Social interaction in the CATS group recommender. In: Workshop on the Social Navigation and Community Based Adaptation Technologies (2006)
23. McCarthy, K., Salamo, M., Coyle, L., McGinty, L., Smyth, B., Nixon, P.: CATS: A synchronous approach to collaborative group recommendation. In: FLAIRS Conference, vol. 2006, pp. 86–91 (2006)
24. McGrath, J.E., Berdahl, J.L.: Groups, technology, and time. In: Scott Tindale, R., et al. (eds.) Theory and Research on Small Groups, pp. 205–228. Springer, US (2002)
25. Nunamaker Jr., J.F., Briggs, R.O., Mittleman, D.D., Vogel, D.R., Balthazard, P.A.: Lessons from a dozen years of group support systems research: A discussion of lab and field findings. *J. Manage. Inf. Syst.* **13**, 163–207 (1996)
26. O'connor, M., Cosley, D., Konstan, J.A., Riedl, J.: Polylens: a recommender system for groups of users. In: Prinz, W., et al. (eds.) ECSCW 2001, pp. 199–218. Springer, The Netherlands (2001)
27. Pommeranz, A., Broekens, J., Wiggers, P., Brinkman, W.P., Jonker, C.M.: Designing interfaces for explicit preference elicitation: a user-centered investigation of preference representation and elicitation process. *User Model. User-Adap. Inter.* **22**(4–5), 357–397 (2012)
28. Pu, P., Chen, L.: User-involved preference elicitation for product search and recommender systems. *AI Mag.* **29**(4), 93 (2009)
29. Pu, P., Chen, L., Hu, R.: A user-centric evaluation framework for recommender systems. In: Proceedings of the Fifth ACM Conference on Recommender Systems, pp. 157–164. ACM (2011)
30. Ricci, F., Rokach, L., Shapira, B.: Introduction to Recommender Systems Handbook. In: Ricci, F., Rokach, L., Shapira, B., Kantor, P.B. (eds.) Recommender Systems Handbook, pp. 1–35. Springer, US (2010)
31. Saaty, T.L.: Fundamentals of Decision Making and Priority Theory with the Analytic Hierarchy Process, vol. 6. RWS Publications, Pittsburgh (2000)
32. Walther, J.B.: Computer-mediated communication impersonal, interpersonal, and hyper personal interaction. *Commun. Res.* **23**(1), 3–43 (1996)

# The #selfiestation: Design and Use of a Kiosk for Taking Selfies in the Enterprise

Casey Dugan<sup>1()</sup>, Sven Laumer<sup>3</sup>, Thomas Erickson<sup>2</sup>, Wendy Kellogg<sup>2</sup>, and Werner Geyer<sup>1</sup>

<sup>1</sup> IBM Research, Cambridge, MA, USA

{cadugan, werner.geyer}@us.ibm.com

<sup>2</sup> IBM Research, Yorktown, NY, USA

{snowfall, wkellogg}@us.ibm.com

<sup>3</sup> Information Systems and Services, University of Bamberg, Bamberg, Germany  
sven.laumer@uni-bamberg.de

**Abstract.** This paper describes the design and use of the #selfiestation, a kiosk for taking selfies. Deployed in an office of a large enterprise, its use was studied through analysis of 821 photos taken by 336 users over 24 weeks and interviews with 10 users. The findings show high adoption amongst residents (81.5 %); describe selfie usage patterns (funatics, communicators, check-ins, doppelgangers, and groupies); illustrate social photo-taking behavior (78.6 % of users posed as part of groups, and those who did took almost four times as many photos); and raises questions for future investigations into flexibility in self-representation over time. Office residents seeing social and community-building value in selfies suggests that they have a place in the enterprise.

**Keywords:** Selfies · Faces · Social media · Enterprise · Self-representation

## 1 Introduction

Selfies, or photographs taken of oneself, have invaded popular culture. But the desire to capture photographs of oneself is not new. For example, an early account of using a photobooth at a conference was published in 1989 [8]. However, traditional photobooths didn't benefit from the design patterns commonly found on social media sites today. Social media sites, from MySpace to Twitter, have made it fast and easy to create an online identity, post content over time, offer network mechanisms among users, and distribute content using these networks. Analysis of Instagram identified self-portraits and portraits with friends among the most popular of eight categories of photos [5]. The abundance of such photos and use of the term "selfie" offers researchers a new opportunity to easily find and create a corpus for studying photographs of humans. Prior research in this area has found that Instagram photos with human faces are 38 % more likely to receive likes and 32 % more likely to receive comments, regardless of age and gender [1]. Other researchers have analyzed images

of groups of people as they are captured in many social settings (overview in [4]). Yet, to date, there has been very few studies analyzing selfies, with current research focusing on interfaces for taking selfies, such as tools to help pose for better selfies [11] and interactions to trigger the photos [6]. Exceptions to this include recent work on the Moment Machine, a public kiosk for taking photos, that's usage was studied over 12 weeks [7] and analysis by those at Selfiecity, who studied 3,200 Instagram selfies from 5 cities around the world [9]. The current popularity of the “selfie” phenomenon, complex issues around presentation of self and perception by others, and their applicability to broader research, warrants further study of this kind of photo.

Early in 2014 we built the #selfiestation, a kiosk for users to take photos of themselves, and deployed it in an IBM office. As the kiosk displays photos taken over time in the physical space, which has a predefined “network” of residents and visitors, its setup is comparable to social media sites in many ways. IBM also has a history of adopting social media tools, such as blogs (2002), social bookmarking (2004), and social networking (2007) [3] and we were motivated to understand how employees would engage with this new form of social media. What began as an exploration into keeping a visual history of visitors and residents, evolved into a long-term observational study of the creation and use of selfies and a first account on the use of selfies in the workplace. This study draws on usage data (821 photos of 336 users over 24 weeks) and 10 user interviews. Our analysis addresses the overall adoption in the workplace, motivations for its use and implications for the work environment, as well as patterns evolving in its use. This joins a body of prior research on social media tools in the workplace, such as enterprise social networking [3], which have shown employees use these tools for search, sharing, discovery of information, and connecting with colleagues. Research has also identified photo sharing as a stimulus for conversations and collaboration, which might also provide the base for the maintenance of social presence among group members [10].



**Fig. 1:** The #selfiestation application

## 2 System

The #selfiestation (Fig. 1) is a full-screen application running on a “55” touchscreen, with a webcam attached to the top of the screen. The screen is divided such that the left side is a live feed “preview” of the webcam, that is not captured or analyzed unless the user presses a large “take my picture” button at the bottom of this section. A notice at the bottom of the screen informs users that by saving an image they acknowledge it can be used for research purposes. The right side of the screen displays photos taken at the kiosk, ordered by recency (most recent at the top).

After pressing the “take my picture” button, a 3-s countdown appears, after which the live preview is frozen and a popup is presented to indicate the name and affiliation for each person in the photo, and a caption. To aid users in filling in these optional fields, the system runs face recognition on the photo and presents recommendations for past visitors the system believes may be present. Additionally, both the name and affiliation fields have auto-complete functionality, which show the time of last visit and potentially a thumbnail of the person to help with disambiguation.

After saving this information, users are shown the photo taken and a ‘Recognition Section’ to welcome new visitors or acknowledge repeat visitors. This has a conditional UI with over 20 possible combinations, based on the number of people listed, whether a repeat visitor is present and how long it has been since the last visit, the affiliations of those in the photo, etc. A repeat visitor is shown their prior selfies and told they were missed if it has been more than 24 h or “you were just here!” if they visited minutes ago, while a new visitor might be shown photos of others with that affiliation. A group is told that the #selfiestation loves group-selfies and how many people they need to hold a record for most people in a photo. If no people are listed, users are told “I love pictures of walls! Maybe next time you’ll list someone!”

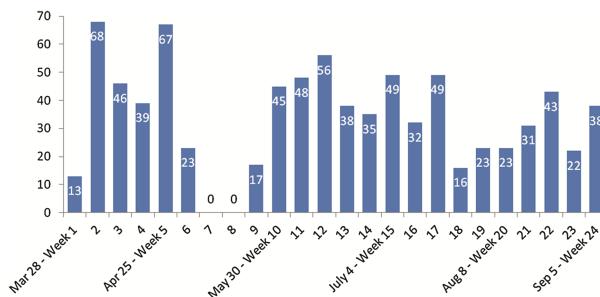
## 3 Results

The #selfiestation was deployed in an IBM office located in Cambridge, MA, U.S. on March 21<sup>st</sup>, 2014. The space has 85 residents, from 3 divisions: Research, Consulting, and those who run a briefing center (Innovation Center), with the latter two groups moving in 5 months before. The space requires badge-entry, limiting access of non-residents (other IBM employees & external guests). We analyzed 24 weeks of usage (launch – September 11<sup>th</sup>, 2014). Photos of researchers were removed, though group photos with non-team members were included. In May, interviews were conducted with 10 users, all IBMers, with a range of ages (20 s–60 +), time with IBM (1yr–20 +), usage (1 photo–20+), males (6)/females (4), and residents (8)/non-residents (2).

### 3.1 #selfiestation: Photos

During the 24-week study period, 821 photos were taken. Sustained usage is seen throughout this time (Fig. 2), though an initial “novelty” effect seems to occur when it was first deployed and when face recognition was introduced (peaks in weeks 2/5).

The kiosk was unavailable (for unrelated reasons) from May 6<sup>th</sup>–May 29<sup>th</sup> (no photos weeks 7/8, dip in weeks 6/9). Coincidentally, interviews were conducted during this time and its absence was commented on in 6 of the 10–3 commented on it being down (e.g. “And actually, today... it’s not set up in #selfiestation mode as I’ve seen it on other days.”) and 2 mentioned it prevented their use (e.g. “I enjoy looking at their selfies, so I was going to go there. Then I remembered [it is down].”). One user went as far as: “I don’t know about other people you’ve talked to, but I definitely miss it.”



**Fig. 2.** Selfies taken over the 24-week study period.

As expected in an office environment, very few photos were taken on weekends (10 total). A one-way between subjects ANOVA showed a significant effect ( $p < .05$ ) between the day of the week and the number of selfies taken ( $F(16,161) = 13.61$ ,  $p = 0.00$ ). Post hoc comparisons using the Tukey HSD test indicate the number of selfies taken on either Friday ( $M = 9.75$ ,  $SD = 7.86$ ) or Thursday ( $M = 7.17$ ,  $SD = 5.61$ ) was significantly higher than the number of selfies taken on any other day of the week (Monday  $M = 4.41$ ,  $SD = 3.95$ ; Tuesday  $M = 5.29$ ,  $SD = 3.95$ ; Wednesday  $M = 5.29$ ,  $SD = 5.13$ ; Saturday  $M = .29$ ,  $SD = .11$ ; Sunday  $M = .08$ ,  $SD = 0.57$ ).

### 3.2 #selfiestation: Individuals and Groups

Researchers identified 336 unique people who posed at the #selfiestation. The number of photos taken by users follows a long-tail distribution commonly seen for user contributions. Users took 3.88 photos on average, with the majority (57.7 %) taking one photo, and a few outliers taking a large number (81, 67, 64, etc.).

Of the 821 photos taken, the majority (60.9 %) contained exactly 1 identifiable human face (500). 33.0 % (271) were group photos containing 2+ people (3.29 people on average, max 24). 6.1 % of the photos (50) contained 0 identifiable faces. These included photos of food, the wall in front of the kiosk, and the kiosk itself.

The physical location of the kiosk allowed us to measure participation among residents. Of the 81 residents (85 minus 4 researchers), 66 took one or more photos, a participation rate of 81.5 %. Separating residents by IBM division, shows participation rates are high across divisions: Consulting (25 of 30 residents, 83.3 %), Research (21 of 22, 95.5 %), Innovation Center (14 of 21, 66.7 %), other divisions (6 of 8, 75.0 %). Residents make up 19.6 % of total users. While we are unable to gauge participation for

the 270 visitors, an independent-samples t-test shows residents take significantly more photos ( $M = 12.77$ ,  $SD = 16.89$ ) than visitors ( $M = 1.71$ ,  $SD = 1.53$ );  $t(65) = 5.313$ ,  $p = 0.00$ . This is expected, as residents have more opportunities to use the kiosk. The 270 visitors included other IBMers (78), guests from external affiliations (89 with 26 different affiliations), family members of residents (discussed in Usage Pattern section), and those who might be overlooked in an office space, such as the person who cleans the office at night. One user said: “there’s an adorable picture of the guys that were putting up the wallpaper...They were not even employees. They were contractors who engaged with the system because they probably saw other people engaging with it.” Interestingly, neither the contractors nor cleaner left their names.

We analyzed the use of the name field to further understand how people represented their identity. A large number did not leave their name on one or more occasion, which could point to UI flaws or users being too lazy to fill it in (we noted each person was rarely listed in photos with a large number of people). However, issues of anonymity is apparent in at least one case, where a resident took 19 photos over many months and never left his name. In cases where he posed with a group, a group moniker was sometimes used rather than labeling individuals in it. Group identities, consisting of a name being used to represent an (often) consistent group of people over time, seemed to form. One example was “burrito boys” consisting of 2 boys and 1 girl who used this name in photos when they had burritos for lunch, not in other photos together (for which they sometimes used other group identities). There were also cases of individuals using different names over time, such as one who used “Johnny” in most photos, but also “John” and “John [Anonymized] Jingleheimer Schmidt” and another who first used “KATE” then switched to “katee” (her last initial was E).

We also coded faces by gender. More men than women visited the #selfiestation: 54.1 % vs 45.8 %. However, this imbalance could be due to more men walking through the space. To better compare, we looked at average number taken by men ( $M = 4.38$ ,  $SD = 5.16$ ) versus women ( $M = 3.29$ ,  $SD = 10.90$ ), but this difference was not statistically significant (independent-samples t-test;  $t(267) = 1.12$ ,  $p = 0.230$ ).

### **3.3 #selfiestation: Motivations and Usage Patterns**

To explain the high participation amongst residents, we turned to motivations given in interviews. The most common reason for using the kiosk (all 10 interviews) was to have fun: “fun just to take a picture of yourself.” Some also received social acknowledgement: “I’ve seen other people walk by other people and say ‘hey, I liked your cool selfie in the station,’” with another saying “it made me feel good” when others found his photo “hilarious.” We identified happiness and increased job satisfaction: “it makes me happy,” “everyone else is having fun with it. And it felt good to get to know colleagues that way.” Community building was cited: “bringing work and people, something that’s more personal, together” and “it would prompt me to approach them more...It captures them in a good moment and they seem very approachable.”

Increased awareness of others was mentioned: “More aware definitely...I’ll see some selfies of people and I’ll be like ‘Hmm, I think I’ve seen that person, but I’ve never spoken to them.’” As described, various groups were brought together and

learning the names of others was mentioned: “I’ve seen people’s names that I didn’t know before,” “There are a bunch of people on our floor who I don’t know... It is a good way to learn who other people on the floor are.” Nearly all said the kiosk was impossible to miss – because of the location in the space (“I think the fact that it is physically there that I’m always reminded of it,” “It’s easy for me to just glance over and take a peek”) or the photos displayed (“The right side has pictures of people that have done it already. So it’s kind of like a mosaic of people. That’s what I found intriguing”). Another said the live camera feed drew his attention. Every resident interviewed said they looked at least once a day, some multiple times daily. There was a belief it would continue. One user said, due to Innovation Center traffic, “The population coming in and out will be changing a lot...there will always be a reason to consult the #selfiestation, to see who has been around and who is taking selfies.”

Based on described motivations & observed behavior, we identified different usage patterns: Funatics, Check-ins, Communicators, Doppelgangers, and Groupies.

The **Funatics** pattern describes taking intentionally funny, playful, or creative selfies hoping to engage others, such as through memes. One meme involved different pairs of eye-glasses held up, all with the name “The Lens” (6 photos over 3 days) and 2 close-ups of eyes, 1 with the caption ‘contact lens.’ The “ghost” meme began with 2 photos of no people, labeled with the name “ghost” (June 9<sup>th</sup>, July 11<sup>th</sup>). Next, the ghost appeared with 2 Innovation Center interns (July 14<sup>th</sup> & 15<sup>th</sup>), then a 3<sup>rd</sup> Innovation Center intern joined them (August 1<sup>st</sup>, their last day), followed by the ghost jumping into a photo with 3 Consulting residents (August 1<sup>st</sup>). The ghost first appearing alone, then with groups of increasing size, might point to the presence of a group emboldening our photo behavior. It also shows such behavior can act as a contagion, transferring from one group to another. Another Funatic behavior was photobombs, or intentionally jumping into another’s photo; we counted 31 photobombs.

The **Check-in** pattern describes photos taken to show that someone is arriving or leaving the office, especially evident in photos containing one person. One user interviewed described the #selfiestation as “like a voluntary census poll.” Typically in these photos, the user leaves a caption like “late” on the way in and a future destination on the way out, e.g. “off to Yorktown.” Often, the user will face the entrance or exit, depending on their direction, rather than face the camera. A series of 6 photos depicting this pattern, taken one night from 9 pm–7 am, chronicled an employee’s overnight stay in the office working – in this case “not leaving” was the check-in.

The **Communicator** pattern describes photos left hoping to communicate something to others in the space. This included the Senior Vice President of the Consulting division leaving a message for residents in her division in the form of a caption addressing them. The interviews were especially helpful in identifying Communicator examples, such as one user saying that he sometimes posed pointing to his baseball cap or tagged them with “go-sox”, and described this as “try[ing] to generate fan support for the Boston Red Sox.” Another described taking a photo with a poster of a speaker the office would be hosting, to notify other residents of the talk.

The **Doppelganger** pattern describes the curious practice we observed in 10 photos where users pose with a likeness of themselves. The likenesses observed included

badges, old driver's licenses, images on a phone or tablet, and a puppet. Sometimes the person labeled themselves twice to indicate they appeared more than once.

Finally, the **Groupie** pattern describes those whose #selfiestation identity is closely tied to the groups they are posing with or with a special interest in publicizing their network. While the majority of the photos taken contained only individuals (60.9 %), analyzing from a user-perspective tells a different story. Of the 336 users, the majority took at least one photo with others (78.6 %). Further, 190 (56.5 %) users \*only\* appeared in group photos. Those interviewed spoke of preferring to pose in a group: "for me, I'm kind of one of these types that likes to do it with people. So my selfies will be when I'm with someone" and "I guess I never stopped by myself and have a selfie taken" as one user said who seemed to feel a kind of peer-pressure to take a selfie with the group he was with. The existence of such a large amount of users whose entire experience was defined by the groups they posed with led us to further analyze group behavior. We found that those who had ever posed as part of a group take significantly more photos on average ( $M = 4.63$ ,  $SD = 9.74$ ) than those who only pose by themselves ( $M = 1.17$ ,  $SD = 0.50$ ):  $t(268) = 5.737$ ,  $p = 0.00$ .

Some users seemed to publicize being part of a group. For example, 20 photos show residents posing with family (13 residents, 26 family members such as children). As a photo of an office visit meant for family could have been taken on a camera phone, we believe it was to show their family to their colleagues. Those interviewed said group photos gave them insights into others' networks: "But I saw them in the #selfiestation and I was immediately able to think 'oh, I didn't know these people knew each other. And that they interact together.'" The strength of ties was also speculated on, "They must know each other pretty well to take a selfie together....if I met some new people, I wouldn't be like 'hey, let's go take a selfie together!'" Of the 147 resident group photos, 83.7 % show residents from the same division, with only 24 from mixed divisions, also suggesting group photos may indicate tie strength.

## 4 Discussion and Conclusion

We deployed the #selfiestation, a kiosk allowing users to take photographs of themselves, in an office location of IBM. Our contributions include a longitudinal study (6 months) of usage patterns as well as the effects of such a kiosk in the workplace. We analyzed the 821 photos taken by 336 users and conducted 10 user interviews.

Our findings show the #selfiestation experienced high participation among residents (81.5 %). The number one motivation mentioned was to have fun. Interestingly, #selfiestation usage increased at the end of the work week – perhaps employees felt this was the most acceptable time for having fun and taking selfies at work. Users described how the kiosk increased their happiness at work, humanized colleagues, and increased community building, with some mentioning the positive feedback they got from others on their photos and saying they would be more likely to reach out to those whose photos they had seen at the kiosk. This ties in with past research on the value of social networking at work [3] and the ability to learn about office culture through photos [10]. Due to the perceived benefits in the workplace, we have begun deploying kiosks at other IBM office locations around the world (Shanghai, São Paulo, etc.).

In this office, employees from various divisions had recently moved in and users mentioned the benefit of learning others' names. This suggests that other spaces with people who may not know each others names, such as conferences, may benefit from such a kiosk. As our kiosk was in a public space, we didn't require users to authenticate and they had flexibility in choosing names to label photos. Preliminary findings show some users made use of this, including using different names over time, group identities, etc. Studying the motivation behind this and the perception by others would give further insights into the conflicts between users and social media sites requiring "legal" names, such as GooglePlus & Facebook [2]. The trade-off of representational flexibility versus benefits of learning others' names requires further study.

We identified 5 #selfiestation usage patterns: Funatics, Check-ins, Communicators, Doppelgangers, and Groupies. The #selfiestation was designed to be a social experience, with half the screen devoted to seeing photos taken by others and the ability to leave a caption. Users took the opportunity to engage with others, clearly seen in the Communicator pattern and Funatics pattern memes. Group behavior was strong, with 78.6 % of users posing with others at least once, and the majority (56.5 %) \*only\* posing in groups. Those interviewed felt the group photos represented a social network that was sometimes previously invisible to them and speculated ties were especially strong among those posing together. Further research could study these networks and strength of ties, and compare it to the company's social networking platform. We also found that those who had ever posed as part of a group posted four times as many photos as users who did not pose with others. Posing with a group could have reduced the barrier to participation, also consistent with our interviews.

Further work is needed to see if these patterns translate to other social media sites, such as Instagram. For example, the nature of the physical kiosk may have made it easier for users to discover and participate in memes. It is clear some behaviors translate from camera phone selfies to the #selfiestation, such as photobombing. Our research is also consistent with prior research on the engaging nature of human faces [1] – with 93.9 % of our photos containing faces and users mentioning that seeing these was compelling, with residents looking daily, some multiple times a day.

## References

1. Bakhshi, S., Shamma, D.A., Gilbert, E.: Faces engage us: photos with faces attract more likes and comments on Instagram. In: Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems. ACM (2014)
2. Boyd, D.: The politics of "real names": power, context, and control in networked publics. Commun. ACM **55**(8), 29–31 (2012)
3. DiMicco, J., Millen, D.R., Geyer, W., Dugan, C., Brownholtz, B., Muller, M.: Motivations for social networking at work. In: Proceedings of CSCW 2008, pp. 711–720. ACM, New York (2008)
4. Gallagher, A., Chen, T.: Understanding images of groups of people. In: Computer Vision and Pattern Recognition. CVPR 2009 (2009)
5. Hu, Y., Manikonda, L., Kambhampati, S.: What we instagram: a first analysis of instagram photo content and user types. In: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM 2014) (2014)

6. Jain, A., Maguluri, S., Shukla, P., Vijay, P., Sorathia, K.: Exploring Tangible Interactions for Capturing Self Photographs. In: Proceeding of India HCI 2014, p. 116. ACM (2014)
7. Memarovic, N., Fatah gen Schieck, A., Schnädelbach, H., Kostopoulou, E., North, S., Ye, L.: Capture the moment: “in the wild” longitudinal case study of situated snapshots captured through an urban screen in a community setting. In: The 18th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW 2015), Vancouver, Canada (2015)
8. Salomon, G.B.: Designing casual-user hypertext: the CHI 1989 InfoBooth. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 451–458 (1990)
9. Selfiecity. <http://selfiecity.net>
10. Thom-Santelli, J., Millen, D.R.: Learning by seeing: photo viewing in the workplace. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM (2009)
11. Yeh, M., Lin, H.: Virtual portraitist: aesthetic evaluation of selfies based on angle. In: Proceedings of MM 2014, pp. 221–224. ACM (2014)

# The LuminUs: Providing Musicians with Visual Feedback on the Gaze and Body Motion of Their Co-performers

Evan Morgan<sup>(✉)</sup>, Hatice Gunes, and Nick Bryan-Kinns

School of Electronic Engineering and Computer Science,  
Queen Mary University of London, Mile End Road, London E1 4NS, UK  
`{e.l.morgan,h.gunes,n.bryan-kinns}@qmul.ac.uk`

**Abstract.** This paper describes the LuminUs - a device that we designed in order to explore how new technologies could influence the inter-personal aspects of co-present musical collaborations. The LuminUs uses eye-tracking headsets and small wireless accelerometers to measure the gaze and body motion of each musician. A small light display then provides visual feedback to each musician, based either on the gaze or the body motion of their co-performer. We carried out an experiment with 15 pairs of music students in order to investigate how the LuminUs would influence their musical interactions. Preliminary results suggest that visual feedback provided by the LuminUs led to significantly increased glancing between the two musicians, whilst motion based feedback appeared to lead to a decrease in body motion for both participants.

**Keywords:** Musical interaction · Computer-supported cooperative work · Groupware · Eye-tracking · Social signals · Non-verbal communication

## 1 Introduction

There has been much focus recently on the use of wearable sensors for lifestyle, sports and activity tracking. Our research concerns the use of wearable devices as tools for understanding and enhancing interactions between musicians during co-present performances. In a previous study we assessed the suitability of various sensors for providing continuous information on the affective and behavioural aspects of collaborative music making [1]. Following on from this work, we decided to design and test a device that would enable musicians to receive real-time sensor-based feedback on aspects of their collaborative interactions. We chose to focus specifically on two non-verbal interaction features that have important expressive and communicative roles in musical performance - *eye gaze* and *body motion*.

Studies have shown that gaze has a variety of functions in non-verbal communication, including the expression of emotional and interpersonal information such as liking, attentiveness, competence, attraction, and dominance [2,3].

It may also be used as a means of directing attention towards shared references and surveying activity in the external environment [4]. Schutz notes that visual co-presence during collaborative music making allows each performer to share “in vivid present the Other’s stream of consciousness in immediacy” [5, pp. 176]. Despite its apparent significance, surprisingly few studies have been conducted on the use of gaze during collaborative music making. Davidson and Good [6] observed that members of a string quartet used “conversations with the eyes” to convey important information while playing. A study with a string duet found that visual contact had a positive influence on timing synchrony between musicians for certain types of music [7].

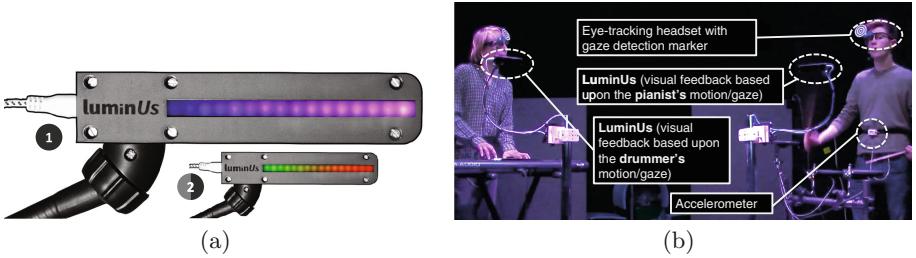
Eye-tracking technology makes it possible to automatically track where someone is looking. Existing studies have predominantly investigated how eye-tracking might be used to support and evaluate the success of computer-supported collaborative work [8,9]. Music related studies have generally focused on the potential for eye-tracking to be used as a control medium (e.g. selecting notes by glancing at specific markers - for a review see [10]). We are not aware of any studies that have used eye-tracking to investigate or support the interpersonal and affective functions of gaze during collaborative musical interactions.

As an intrinsic feature of non-verbal communication, body motions are closely tied to gestural [11] and emotional expression [2]. Humans are able to successfully identify emotions from simple point-light representations of body motion [12,13]. More specifically, studies have found evidence for a link between measures of body motion activity and the arousal/activation aspects of emotion [14,15]. Body motion analysis has also been applied to the study of affect in musicians [16]. Motion measurements can be obtained using accelerometers worn at positions of interest. Many consumer electronic devices - such as smart phones, smart watches, and activity trackers - already contain accelerometers, from which data can be accessed using custom built applications.

In this paper we describe the design of the LuminUs - a device that provides musicians with real-time visual feedback on either the gaze or body motion of their co-performers. We also report preliminary results from a study that investigated the impact that the LuminUs has on dyadic improvised performances.

## 2 The LuminUs

We designed the LuminUs with the aim of exploring how technology might be used to provide performing musicians with an increased awareness of each other, especially in situations where their mutual attention might be hindered by complex musical interfaces and physical obstructions. The LuminUs consists of a strip of 16 coloured LEDs that can be controlled individually to provide dynamic visual information (see Fig. 1(a)). The device is mounted on a flexible arm so that it can be positioned in front of the musician. Control of the lights is provided via USB connection to a computer. The design was chosen based upon three criteria for the way in which feedback should be presented to the musician: *visual* (as opposed to haptic or audio); *minimal* (so as not to be overly distracting); and *dynamic* (to represent time varying signals).



**Fig. 1.** (a) The LuminUs: different colours were used for gaze (1) and motion (2) feedback; (b) Annotated image of the experimental setup.

The LuminUs has two modes of operation - *motion feedback* and *gaze feedback*. In the motion feedback mode we measure the body movements of each musician using small wireless accelerometers worn around the waist. These movements are then processed and displayed on the LuminUs. The greater the movement, the more lights are illuminated - with the colours of the lights ranging from green (low) to red (high). In the gaze feedback mode we use eye-tracking glasses to detect when one musician is glancing towards the other. The LuminUs then visually notifies the musician who is being glanced at. In this case, more lights become illuminated as the duration of the glance increases. For gaze feedback the colours of the lights range from blue (short) to purple (long).

For eye-tracking we used the Pupil headset, which is an open-source hardware and software project [17]. The headset tracks the movement of the right eye using a single infra red camera, and simultaneously captures the wearer's point-of-view (POV) using a separate forward-facing video camera. Following a short calibration ( $\sim 20$  s), the software is able to map the gaze point of the wearer onto the live POV video. We modified the pupil software so that it could recognise 2D markers within the POV images. Consequently, by placing markers on the headsets of each musician (see Fig. 1(b)), we are able to automatically register the moments when one person is glancing at the other.

### 3 The Experiment

Our hypotheses concerning the way in which the LuminUs would influence musical interactions were as follows:

- H1.** Providing gaze feedback would increase the overall amount of gaze during collaborative interaction.
- H2.** Providing motion feedback would stimulate increased awareness of the other participant, as indicated through more glancing at the other.
- H3.** Providing motion feedback would increase the overall amount of motion during collaborative interaction.

To test these hypotheses we designed an experiment in which pairs of pianists and percussionists were asked to create live improvised accompaniments to a silent

animation. This task was chosen for a number of reasons. Firstly, the improvisational nature of the task encouraged the musicians to work collaboratively and heightened the need for mutual awareness. Secondly, the animation meant that the musicians had more to attend to than simply their instrument and co-performer - as would often be the case in real-world performances. In addition to this, it provided a visual stimulus to guide and influence the improvisations created by each pair.

Four two minute long animations were short-listed for use in the study. We then asked a separate group of 24 musicians to watch and rate the animations based upon their suitability for improvised musical accompaniment. The animation that received the highest average rating was chosen for use in the study.

### 3.1 Method

15 percussionists and 15 pianists were recruited from music colleges in London. The participants comprised 7 females and 23 males aged between 18 and 38 ( $M = 22.7$ ,  $SD = 4.7$ ). Their playing experience ranged from 3 to 33 years ( $M = 11.6$ ,  $SD = 5.6$ ). The participants were assigned to percussionist-pianist pairs such that the individuals in each pair did not know one another.

There were seven experimental conditions, coded as: G-G, G-X, X-G, M-M, M-X, X-M, and X-X. The first letter represents the feedback that the LuminUs provided to the percussionist, whilst the second letter represents the feedback provided to the pianist - either gaze feedback (G), motion feedback (M), or no feedback (X).

The experiment was held in a performance space, with overhead stage lighting and black curtains surrounding the performance area. The setup consisted of a large screen, which showed the animation and provided instructions to the participants. The two participants were then positioned facing the screen, but angled slightly towards one another. The percussionist was provided with two electronic drum pads (snare and floor tom) and an electronic ride cymbal, which they played standing up. The pianist was provided with a 61 note keyboard and sustain pedal, which they also played standing up. Both instruments were connected to a computer via MIDI, and the audio was output through speakers positioned either side of the screen. Each participant's LuminUs was positioned on a stand in front of them, such that it was just below their line of sight to the screen. The devices were positioned such that each participant could only see their own device, and the brightness of the lights was adjusted so that no reflections were visible. Figure 1(b) shows the experimental setup as seen from below the screen.

Upon arriving for the experiment the participants were given a couple of minutes to play the instruments together. This gave them a chance to make brief introductions and familiarise themselves with the instrumental set up. The participants were then fitted with the accelerometers and eye tracking headsets, which were subsequently calibrated for each participant. All of the eye-tracking and motion data was time-stamped and saved. The experiment began with a warm up session, where participants watched the animation twice without playing their instruments and then twice whilst playing along. Following

this, the participants were given roughly one minute to discuss ideas. This was the last opportunity for them to verbally interact during the experiment. The remainder of the experiment consisted of seven improvisation sessions, each randomly assigned to one of the seven conditions discussed above. Prior to each session each participant's LuminUs would show either green-red (motion feedback), blue-purple (gaze feedback), or no lights (no feedback) to indicate which type of feedback they would receive in that session. This was done so that neither participant would know what kind of feedback the other was receiving. In each session the participants had two attempts to play along to the animation, with a 10 s gap between. After each session the participants completed a short questionnaire to rate aspects of their experience and subjective opinions relating to that particular session. The analysis of our questionnaire results is out of the scope of this paper.

### 3.2 Results and Analyses

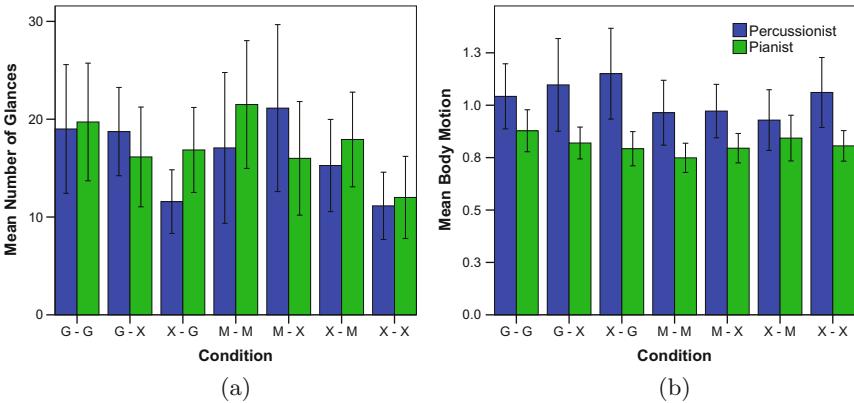
To test **H1** and **H2** we decided to look at the average number of glances that each participant made towards their co-performer within each condition. This information was extracted from our eye-tracking data. Figure 2(a) shows the mean and standard error for the number of glances averaged over participants within each condition. We see that the mean values for both the pianist and percussionist are lowest in the conditions where they are not receiving any feedback from the LuminUs. We can also see that, on average, the pianist tended to glance more than the percussionist.

To test **H3** we extracted the mean body motion for each participant within each condition. The results are shown in Fig. 2(b). We can see that the percussionist moved more than the pianist, as might be expected. More interestingly, it appears that the body motion values tend to be lower when the participants are receiving motion feedback, compared to gaze feedback.

**Table 1.** Statistical results for the differences between means in conditions 1–6 (LuminUs in use), relative to those in condition 7 (LuminUs not in use).

Condition	Number of glances				Body motion			
	Z	r	n	p	Z	r	n	p
(1) G - G	3.81***	0.70	27	.000	1.00	0.18	30	.318
(2) G - X	3.14**	0.57	28	.002	0.30	0.05	30	.766
(3) X - G	2.46*	0.45	27	.013	0.24	0.04	30	.813
(4) M - M	2.83**	0.52	28	.005	-1.61	-0.29	30	.106
(5) M - X	2.43*	0.44	28	.015	-0.87	-0.16	30	.382
(6) X - M	2.61**	0.48	28	.009	-1.14	-0.21	30	.254

Z = Wilcoxon signed rank z-statistic, r = effect size, n = sample size, p = two-tailed p-value. \* $p < .05$ , \*\* $p < .01$ , \*\*\* $p < .001$ .



**Fig. 2.** Bar graphs showing (a) the mean number of glances; and (b) the mean body motion within each of the seven conditions. G = gaze feedback, M = motion feedback, X = no feedback. Error bars:  $\pm 1$  SE.

We used the Wilcoxon signed rank test to perform statistical analyses of some of the relationships in Fig. 2(a) and (b). In particular, we looked at the differences between the mean values obtained in the 6 conditions where the LuminUs was active, relative to the equivalent means for the inactive LuminUs condition (X - X). The results are shown in Table 1. The results for the mean number of glances indicate that there were significant differences between all of the conditions and the X - X condition. This effect is greatest for condition 1, where both participants had gaze feedback ( $r = 0.7, p = .000$ ). The second strongest effect is for condition 2, where only the percussionist had gaze feedback ( $r = 0.57, p = .002$ ). The third strongest is for condition 4, where both participants had motion feedback ( $r = 0.52, p = .005$ ).

For the body motion results we see that none of the differences are significant. However, of potential interest is the fact that a mild negative effect is observed for condition 4 ( $r = -0.29, p = .106$ ), where both participants had motion feedback. In addition to this, the only other weak negative effects correspond to the other conditions involving motion feedback. For the conditions involving gaze feedback (1–3) the effect sizes are small (<0.2).

### 3.3 Discussion

**H1** and **H2** are supported by our findings, which indicate that when participants were receiving feedback from the LuminUs they glanced towards each other significantly more than when no feedback was provided. The size of this effect was greatest for the condition where both participants had gaze feedback. However, it is possible that the main effects were simply due to the LuminUs serving as a reminder of the other musician's presence. We intend to carry out further analyses to investigate the nature of these effects in more detail. In particular, we will look at how the specific timings of glances relate to the light output of

the LuminUs. This will allow us to see whether musicians responded directly to the feedback, or whether it had a more general influence on their behaviour during the sessions.

Regarding **H3**, our results were less conclusive, but appeared to indicate that motion feedback led to decreased body motion. We envisaged that providing feedback on the motion of a co-performer would, if anything, encourage participants to move more. This hypothesis was partly influenced by studies on behavioural mimicry, which suggest that the actions or emotions displayed by one person can cause congruent behaviour in another person [18]. The fact that an opposite trend was observed is a potentially interesting finding. Unfortunately there is an absence of existing research on the behavioural effects of providing real-time motion feedback, so at this stage we can only speculate as to any causality. A possible explanation is that the musicians felt self-conscious when they knew their movements were being displayed on the LuminUs. However, if this was the case then we might expect a similar effect for the gaze feedback. Furthermore, none of the musicians indicated feelings of self-consciousness when we informally interviewed them after the experiments.

The results presented in this paper indicate that the LuminUs had an influence upon the *objective* aspects of the musicians' behaviours. However, at this stage it is not possible for us to say whether these were detrimental or beneficial to the musical interaction as a whole. Our ongoing analyses will investigate whether the use of the LuminUs influenced the subjective experiences of the musicians; the quality of their coordination; and the musical outcomes of their performances. In order to investigate this we will use the self report data collected from the questionnaires alongside the audio and video recordings of the performances. An additional consideration is that the short duration of the study may not have provided the musicians with sufficient opportunity to familiarise themselves with the LuminUs and make effective use of its feedback. In future work it would be interesting to investigate how musicians appropriate the LuminUs over a more sustained period of usage.

## 4 Conclusion

This paper provides a brief introduction to the LuminUs, and some preliminary results from experiments, where we put it to the test with trained musicians. Our initial findings suggest that providing real-time co-performer gaze and motion feedback can have a measurable impact on aspects of non-verbal communication and behaviour during collaborative musical performances. Our continued analyses will attempt to ascertain whether or not the LuminUs also influenced the *quality* of the interactions between musicians and the accompaniments they created. Furthermore, through detailed analysis of our eye-tracking, motion, and video data, we hope to gain a better understanding of how musicians use non-verbal communication during live performances.

**Acknowledgements.** This research is supported by the Media and Arts Technology Programme, an RCUK Doctoral Training Centre in the Digital Economy.

## References

1. Morgan, E., Gunes, H., Bryan-Kinns, N.: Using affective and behavioural sensors to explore aspects of collaborative music making. *Int. J. Hum.-Comput. Stud.* **82**, 31–47 (2015)
2. Kendon, A.: Some functions of gaze-direction in social interaction. *Acta Psychologica* **26**, 22–63 (1967)
3. Kleinke, C.L.: Gaze and eye contact: a research review. *Psycholog. Bull.* **100**(1), 78–100 (1986)
4. Argyle, M., Graham, J.A.: The central Europe experiment: looking at persons and looking at objects. *Environ. Psychol. Nonverbal Behav.* **1**(1), 6–16 (1976)
5. Schutz, A.: Making music together. Collected papers II (1976)
6. Davidson, J.W., Good, J.M.M.: Social and musical co-ordination between members of a string quartet: an exploratory study. *Psychol. Music* **30**(2), 186–201 (2002)
7. Vera, B., Chew, E., Healey, P.: A study of ensemble performance under restricted line of sight. In: Proceedings of the International Conference on Music Information Retrieval, Curitiba, Brazil (2013)
8. Vertegaal, R.: The GAZE groupware system: mediating joint attention in multi-party communication and collaboration. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI 1999, pp. 294–301. ACM Press, New York (1999)
9. Chanel, G., Bétrancourt, M., Pun, T., Cereghetti, D., Molinari, G.: Assessment of computer-supported collaborative processes using interpersonal physiological and eye-movement coupling. In: Affective Computing and Intelligent Interaction (ACII 2013), Geneva, Switzerland (2013)
10. Hornof, A.J.: The prospects for eye-controlled musical performance. In: International Conference on New Interfaces for Musical Expression (NIME 2014), pp. 461–466 (2014)
11. Goldin-Meadow, S.: Beyond words: the importance of gesture to researchers and learners. *Child Dev.* **71**(1), 231–239 (2000)
12. Walk, R.D., Homan, C.P.: Emotion and dance in dynamic light displays. *Bull. Psychon. Soc.* **22**(5), 437–440 (1984)
13. Clarke, T.J., Bradshaw, M.F., Field, D.T., Hampson, S.E., Rose, D.: The perception of emotion from body movement in point-light displays of interpersonal dialogue. *Perception* **34**(10), 1171–1180 (2005)
14. Castellano, G., Villalba, S.D., Camurri, A.: Recognising human emotions from body movement and gesture dynamics. In: Paiva, A.C.R., Prada, R., Picard, R.W. (eds.) ACII 2007. LNCS, vol. 4738, pp. 71–82. Springer, Heidelberg (2007)
15. Kleinsmith, A., Bianchi-Berthouze, N.: Affective body expression perception and recognition: a survey. *IEEE Trans. Affect. Comput.* **4**(1), 15–33 (2012)
16. Dahl, S., Friberg, A.: Visual perception of expressiveness in Musicians' body movements. *Music Percept.: Interdisc. J.* **24**(5), 433–454 (2007)
17. Kassner, M., Patera, W., Bulling, A.: Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction, April 2014. CoRR [abs/1405.0006](https://arxiv.org/abs/1405.0006)
18. Chartrand, T.L., Lakin, J.L.: The antecedents and consequences of human behavioral mimicry. *Annu. Rev. Psychol.* September 2012

# An Artifact Ecology in a Nutshell: A Distributed Cognition Perspective for Collaboration and Coordination

Christina Vasilou<sup>(✉)</sup>, Andri Ioannou, and Panayiotis Zaphiris

Cyprus Interaction Lab, Cyprus University of Technology, Limassol, Cyprus  
{c.vasilou, andri.i.ioannou, panayiotis.zaphiris}  
@cut.ac.cy

**Abstract.** An artifact ecology is an environment where multiple heterogeneous technologies co-exist and are interlinked as a unified system. To construct effective ecologies of artifacts for collaborative activities we need to acquire deep understanding of the complex interactions and interdependencies between users and tools. Researchers have identified Distributed Cognition (DC) as a powerful tool for understanding these interdependencies. In this study, DC, and particularly the DiCoT framework, were considered ideal for constructing this understanding for four student-groups during collaborative activities in an artifact ecology. Using DiCoT we analysed learners' behaviour and how the artifact ecology supported collaboration and cooperation. The cognitive system was described from three different perspectives - physical layout, information flow and artifacts - which (i) allowed an in-depth understanding of the interactions among learners and tools during collaborative activities and (ii) provided insights on how the affordances of the artifact ecology supported collaboration and coordination.

**Keywords:** Distributed cognition · DiCoT framework · Artifact ecology · Technology-rich workspace · HCI education · Collaboration · Coordination

## 1 Introduction

As technology progresses, ubiquitous computing, once envisioned by Weiser [1], is now partially a reality. This evolving nature of technology has brought new possibilities to the design of technology-rich learning environments for collaborative activities. As tablets and smartphones are blended with personal computers in our everyday lives, we are no longer locked in front of a single screen, at work or during learning activities. These technologies communicate and share information with each other creating their own network, an ecology of artifacts [2, 3]. Further, Loke and Ling [4] explained that these heterogeneous technologies are interlinked as a unified system. In the case of collaborative environments, group members may work together tackling the same problem while also work individually on sub-tasks. Digital and physical artifacts within the artifacts ecology may be used for a variety of tasks while each individual may perform a task differently. Therefore, there are endless possibilities and design considerations for the construction of an artifact ecology for collaborative activities.

Salomon [5] claimed that the design and integration of new technologies in an environment cannot be studied independently of its surroundings. To design effective technology-rich environments we need to acquire a deep understanding of the complex relations and interactions between collaborators and information technologies. Distributed Cognition (DC) considers a collaborative activity taking place across individuals, tools and representations, as a unified cognitive system. In the areas of human-computer interaction (HCI) and computer supported cooperative work (CSCW) DC has been identified as a powerful tool for understanding the interdependencies between users, tools and tasks [6]. The added benefit of examining a complex collaborative system through DC is that it allows researchers to take a step backwards and see the “whole picture”, focusing on interactions and actions central to the coordination of collaborative activities [7]. Such an understanding will allow researchers and practitioners to pin-point where changes should occur or should not occur in the cognitive system as a whole.

Specifically, in this work, DC was considered an ideal framework to disclose the fundamental processes for collaborative activities in a multi-participant, multi-tool environment. In this paper, we present an in-class investigation of four groups of postgraduate students during collaborative learning activities in an artifact ecology over a period of 12 weeks. More precisely, the study adopted the Distributed Cognition of Teamwork (DiCoT) framework of Blandford and Furniss [8] which emerged from the need to develop a methodology for DC analysis [8]. Using DiCoT we analysed learners’ behaviour in order to understand the interactions and interdependencies in the environment, between learners, tools, and the physical architecture. The purpose of the study is to illustrate the utility of DC and DiCoT as a tool for modelling interactions and interdependencies during collaborative learning activities in an artifact ecology. In this context, we showcase learner - learner and leaner-artifacts interactions evident in the workspace and highlight the affordances of the ecology of artifacts that support collaboration and cooperation during collaborative learning activities. The paper concludes with implications for designing technology and technology set-ups for collaborative learning activities in an artifact ecology.

## 2 Theoretical Framework

### 2.1 Distributed Cognition

The evolution of cognitive sciences has brought to the forefront the idea that cognition cannot be bounded inside an individual’s mind [9], but should conjointly consider an individual’s surroundings. DC suggests that cognition must be seen as a more complex mechanism, one that encloses cognitive processes outside one’s mind, such as manipulating external objects, transitioning and transforming information between actors and tools. When these cognitive processes are studied during collaborative human activity we can observe the distribution of cognition from different perspectives: distribution amongst members of the group, distribution across the physical or digital structure of the group workspace. Hollan, Hutchins and Kirsch [10] emphasized the importance of understanding the distribution of cognitive processes when designing effective human computer interactions.

Researchers in HCI and CSCW communities have identified DC as a valid tool to understand the interactions and dependencies amongst participants, technologies and activities [6]. Hutchins and Klausen [11] studied the distribution of cognitive processes among members of a cockpit flight crew. They reviewed the interactions between internal and external representations and the architecture of information propagation in the cognitive system. Through their analysis they could identify patterns in the collaboration and coordination of the cockpit crew. Such understanding is important not only for redesigning existing system designs and practices but also for creating the basis for new technologies. For instance, Nobarany, Haraty and Fisher [12] employed DC to design a collaborative system to facilitate analytics. Researchers identified cognitive processes that could be used to support users' collaboration from the beginning, in order to design the system accordingly.

Based on previous studies, DC can provide a detailed identification of issues with existing work practices and mediating artifacts [13]. In addition, DC allows researchers to highlight what is salient in the design of existing collaborative working systems and practices and indicate aspects that require redesigning. In this work, we focus on understanding classroom interactions during collaborative problem-based learning activities within an artifact ecology. Therefore, DC was considered an appropriate framework for building this understanding and highlighting affordances of the artifact ecology supporting collaboration and cooperation.

## 2.2 Distributed Cognition for Teamwork (DiCoT)

Nevertheless, there is no established methodology towards applying DC to a learning environment. Therefore, in order to build a concrete understanding of our data from a DC perspective, we adopted the Distributed Cognition for Teamwork (DiCoT) methodological framework introduced by Blandford and Furniss [8] for collaborative work. DiCoT framework emerged from the need to develop a methodology for DC analysis. It draws on ideas from Contextual Design [14], but re-orients them towards the DC framework. Compared to DC framework analyzed previously, Contextual Design can be viewed as a structured approach to collect and interpret data from fieldwork to build a product [14]. The "context" aspect highlights the need for in situ and field investigations. DiCoT models were derived from the Contextual Design approach but are re-oriented towards principles that are central to the DC framework. Our decision to use DiCoT was based on the fact that DiCoT combines the theoretical framework of DC and the structure that Contextual Design provides, in order to provide an effective modelling tool to investigate and understand human behavior in a socio-technical environment.

DiCoT methodology encloses 22 principles, 18 of which are loosely classified in three models; physical layout, information flow, and artifacts [15]. More particularly, the physical model relates to the physical organization of collaborative activities and covers all aspects which are associated with a physical layout component [8]. This model focuses on factors that influence the way a system performs at a physical level, such as situation awareness, naturalness, bodily movements. Based on the aim and scope of the researchers, their focus can be differentiated, switching between key participants and primary locations/settings of the system.

On the second level, the information flow model, focuses on the flow of information neglecting the design of the mediating artifact by which information is transmitted [8]. There is a diversity of viewpoints, depending on the depth a researcher may want to examine or the issue that needs to tackle, e.g. focusing on the system as a “black box”, focusing on the actors, or the way information flows and is transformed within the system.

Finally, the third aspect of the DiCoT framework - the artifact model - focuses on the detailed design of individual artifacts that are important within the cognitive system [8]. In particular, the artifact model focused on the artifacts deemed important to the activities and highlights the role of individual artifacts and pinpoints their affordances that support or hinder cognition. Table 1 presents the DiCoT principles classified in three models adopted from [8].

DiCoT methodology was evaluated and validated within a large ambulance call control center [15]. The analysis of a complex system through DiCoT may also support reasoning for both existing and future system designs [15]. Furthermore, DiCoT has been also applied in various research studies under the project CHI + MED for evaluating and improving healthcare technology in collaborative working environments such as the intensive care unit [16, 17]. Such an environment is considered of high complexity due to the strict and multiple interdependencies between nurses, doctors and healthcare technology. Stepping outside the healthcare system, DiCoT has also been useful in understanding and expanding the collaboration and coordination paradigm amongst eXtreme Programming teams [18]. Such teams are highly collaborative and self-structured, breaking down the problem into singular tasks. Through these numerous tasks, they manage to keep and distribute the status of each task, coordinating their team successfully.

### 2.3 Artifact Ecology

An artifact ecology is a space rich in technological tools with which individuals interact. These technologies communicate and share information with each other creating their own network [2, 3]. Further, Loke and Ling [4] explained that these devices interact “with one another, with users, and with Internet” (p. 78). Researchers have utilized the metaphor of “ecology” to indicate the co-habitation of multiple heterogeneous devices that are interlinked, acting as one system. In the case of collaborative environments, group members may work together tackling the same problem and working on sub-tasks individually. Digital and physical artifacts within the artifact ecology may be used for a variety of tasks while each individual may perform a task differently. Therefore, there are endless possibilities and design considerations for the construction of an artifact ecology for collaborative activities.

Researchers from various fields of engineering, design and education, have designed and investigated artifact ecologies from their own perspectives. For example, artifact ecologies have been designed to improve problem solving activities [19], support classroom learning [20], boost creativity in design conversations [21], or orchestrate teamwork in complex systems [22]. All these different artifact ecologies

**Table 1.** DiCoT principles per model

	No	Principle name and description
Physical Layout	1	Space and Cognition: Space as a medium of supporting cognition during an activity.
	2	Perceptual: Spatial representations supporting cognition.
	3	Naturalness: Each representation match the features of that which it represents.
	4	Subtle Bodily Supports: How bodily actions are used to support activity.
	5	Situation Awareness: How are people kept informed of the activity.
	6	Horizon of Observation: What can be seen or heard by a person.
	7	Arrangement of Equipment: Physical arrangement affecting access to information.
Information Flow	8	Information Movement: Mechanisms used to move information.
	9	Information Transformation: How information is transformed in the system.
	10	Information Hub: Central point of information flow and decisions.
	11	Buffering: Hold up information until it can be processed.
	12	Communication Bandwidth: Richness of information during communication.
	13	Informal and Formal Communication: Importance of informal communication channels.
	14	Behavioral Trigger Factors: Individuals act in response to certain behavior.
Artifacts	15	Mediating Artifacts: Elements used to fulfill an activity within the system.
	16	Creating Scaffolding: How people use the environment to support their actions?
	17	Representation - Goal Parity: How close is the representation of current and goal state?
	18	Coordination of Resources: Plans, goals, history etc and their coordination to support cognition.

have been designed each time taking into consideration the people, the activities and the aim of the setting.

When it comes to collaborative learning activities in a real world setting, as team-members work together with a particular goal in mind, several tasks run at the same time and each team member may acquire a different way of performing a task. As Huang et al. [22] highlighted, projections, screens, and interactive displays have clear interdependencies within an ecology, although not designed as a unified system.

Laying technologies next to each other to work together will result into a wider cognitive system. To design effective collaborative learning environments we need to acquire a deep understanding of the complex relations and interactions between collaborators and information technologies. Looi, Wong and Song [23] stressed the importance of what affordances or constrains different technologies such as mobile devices can bring to a technology-rich environment. More recently, surface computing technologies have been embedded in classroom settings creating an artifact ecology to support collaboration [24]. What is important, however, is to understand what each one of these technologies brings to the collaboration and coordination of group-work.

Based on previous studies in technology-rich environments, DC can provide a detailed identification of issues with existing work practices and mediating artifacts [6, 7, 13]. In addition, DC is prompted on highlighting what is salient in existing collaborative working system designs and practices and indicate aspects that require redesigning. The added benefit of examining a complex system through the lens of DC is that it allows researchers to take a step backwards and see the “big picture”, focusing on interactions and actions central to the coordination of work activities [7]. Such an understanding will allow researchers and practitioners to pin-point where changes should occur or should not occur in the cognitive system as a whole. As explained earlier, DiCoT framework explicitly emerged from the need to develop a methodology for DC analysis [8].

In this study, DC and the DiCoT methodological and analytical framework, were considered ideal for constructing this understanding for four student-groups during collaborative learning activities in an artifact ecology.

### 3 Methods

#### 3.1 Participants

Participants in this study were 21 students (13 female) attending a postgraduate HCI course. Students were assigned to groups of four to five members. For the allocation of students in groups, we kept in mind the aim of creating multidisciplinary groups. Thus the procedure of forming groups was in part based on each student’s background, including studies in computer science and games, graphic arts and interactive multi-media, and education and communication media. Therefore, each group was composed of members from different disciplines.

The age span of the participants was between 22–45 years old ( $M = 30$ ). All students were familiar with digital technologies such as smartphones and tablets as well as with social networking spaces.

#### 3.2 Context

The course was organized in three-hour weekly face-to-face sessions and followed a problem based learning (PBL) approach. Every session included phases of problem analysis, research, reporting and reflection as indicated by Koschmann and Stahl [25] while research could also occur online in-between the face-to-face sessions. PBL

enables students to draw the path of their own learning while working within a group towards the solution of an open-ended problem. The tutor provided a short lecture at the beginning of each session in order to provide a triggering point for students' self-directed learning, as it was deemed important by Hmelo-Silver [26]. Students would then get into their groups for the collaborative activities, which focused on providing hands-on experience allowing active collaboration and engagement with the problem at hand.

### 3.3 Design Problem

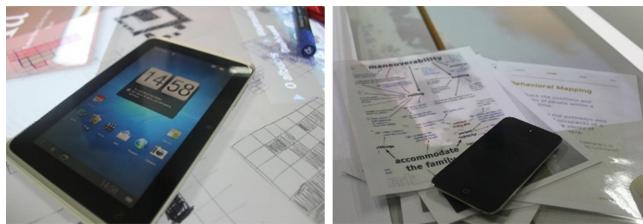
The selection of the problem scenario is a crucial aspect of PBL. For the current in-class investigation the problem was derived from the student design competition of CHI 2007, entitled "Changing the Perspectives of Public Transport" and indicated the need to design an object, product or system that would promote the use of public transportation in Cyprus. The selected problem provided an open and real-world call for action, challenging students to provide a solution that could help drivers and improve the local transport infrastructure.

### 3.4 Artifact Ecology

We sought to create an artifact ecology, by enriching the classroom environment with various technologies aimed to support student collaborative activity, particularly brainstorming, researching, reporting or reflecting, both in-class and in distance (in-between the face-to-face meetings) [27]. Furthermore, this physical space, together with the PBL approach aimed to promote openness and flexibility. Students were encouraged to use the technologies as they perceived appropriate for each activity and task. Each group worked in a physical, technological set-up exhibiting three main attributes that we considered important for collaborative learning activity:



**Fig. 1.** Downward pointing projection



**Fig. 2.** Mobile devices in the artifact ecology

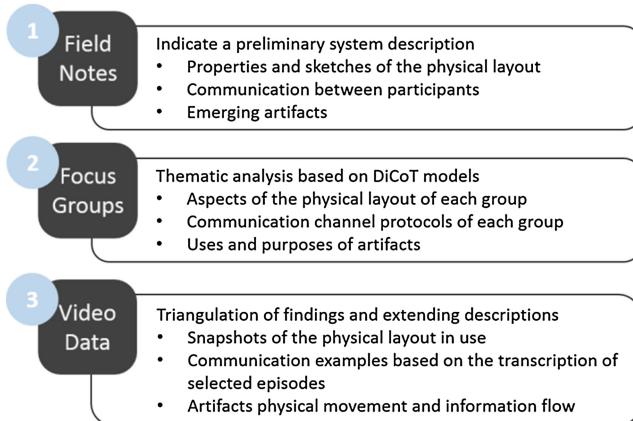
1. A downward pointing projection was provided as a central focus point to support students' fertile discussions and activities (see Fig. 1). This setup would cultivate the blend of physical and digitally projected artifacts, mixing paper and technology, on the same workspace [28, 29].
2. The multitasking nature of the group was invigorated with mobile devices such as tablets, iPods and laptops for concurrent research and record-keeping (see Fig. 2). The students were also allowed to enrich the artifact ecology with their own devices.
3. Last, a widely used social networking platform (Facebook) was used to strengthen information sharing, coordination and collaboration, both between group members and devices.

### 3.5 Data Collection and Analysis

During the three month duration of the study, we observed and kept field notes of weekly sessions regarding group's procedures and the role of technology in their practices. We also collected self-reported data through conducting a focus group with each group at the completion of the course. The focus groups aimed to collect information about students' learning experiences within the artifact ecology, cognitive aspects of their actions and the affordances of the technologies provided. As a triangulation data source, we video-recorded all the collaborative sessions.

Initially, field notes were reviewed to create a preliminary system description guided by concrete principles provided by the DiCoT framework. The preliminary analysis indicated properties of the physical layout, main participants/tools and channels of communication and artifacts emerging as important. Second, we transcribed and reviewed the focus group data, classifying them in three major thematic units according to the DiCoT models - physical layout, information flow and artifacts as in Table 1. From the thematic analysis we retrieved information regarding particular aspects of the physical workspace practices, communication channel protocols and purposes of artifacts' use for each team. Last, we reviewed the video data and selected video episodes that corresponded to principles for each one of the models. Video data assisted us on validating findings from field notes and focus groups (i.e., triangulation of findings). Selected video episodes were transcribed to enrich the existing descriptions of the three DiCoT models constructed from the analysis of the field notes and focus group data. In addition, the review of the video data provided snapshots of the physical layout in real-use, communication channel examples, and artifacts physical movement and

information flow. A schematic representation of the analysis process is demonstrated in Fig. 3 below.



**Fig. 3.** Data analysis procedure

## 4 Findings

The analysis focused on understanding the interactions and interdependencies during collaborative learning activities in an artifact ecology. In this context, we showcase learner-learner and learner-artifact interactions evident in the workspace and highlight the affordances of the ecology of artifacts that support collaboration and cooperation during collaborative learning activities. In the following section, we describe the three DiCoT models, namely physical layout, information flow and artifact that were constructed during the analysis. Each model is described in depth, referring to DiCoT principles and providing additional materials such as information flow examples.

### 4.1 Physical Layout

The physical layout model covers aspects of collaborative learning activities that have a physical layout component. In this case, the group's physical workspace was mainly the downward pointing projection area which was fairly centered in the middle of a table surface leaving a 20 cm wide space around it as individual work area. During PBL activities it was extensively used for research purposes and projecting group artifacts such as working documents or the groups' progress on Facebook (Principle 7). Group members were sitting around the workspace ensuring equal opportunities of access to artifacts and communication channels (Principle 1).

Other digital tools and physical artifacts were spread around the projection, space used as an individual learning space. Laying out their material and working in the individual workspace, learners could still be aware of the group activities (Principle 5).

Furthermore, group members were within each other's zone of normal hearing which resulted in listening to the conversations and issues raised by their group-mates (Principle 6). The focal point and proximity between learners and tools enhanced group awareness, and consequently influenced group's monitoring of events and tacit learning.

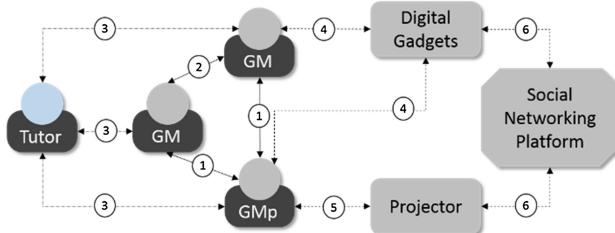
Often, students attempted to interact with the projection naturally using gestures (Principle 3); they falsely perceived the projection as an interactive screen indicating a perceptual shift in human-computer interactions. Furthermore, group members used bodily movements to support their discussion of artifacts projected on the shared workspace, e.g. pointing to screen areas (Principle 4). Pointing directly at the artifact under discussion attracted the group's attention, turning their eye gaze towards the screen, potentially supporting their cognitive processes. For instance, a group observed a video playing and a group member (GM) explained aspects of the video while pointing to the screen.

- GM 1: This is what he is seeing. [Points to a particular area on the screen that encloses the view of the Google Glasses user.]
- GM 2: Yes on the mirror of the glasses.
- GM 1: He tells the Google Glasses to record. Tells it to look for a specific animal on the Internet.
- GM 3: [Points to the screen area where the search results will show up for the user]
- GM 1: It says to capture a photograph.
- GM 4: So you talk and Google Glasses are doing what you instruct them to do?
- GM 1: Yes, yes.

## 4.2 Information Flow

The information flow model pays attention to the perspective that deals with the way information propagates around the cognitive system. In this case, group members were located around the projected surface and freely moved around the downward pointing projection retaining equal opportunities towards the collaborative activities. Yet, the projection was handled by one individual at a time. This restricted the communication channels between individuals and shared projection, since every information had to be channeled towards the group member handling the physical mouse and keyboard operating the projection. The communication between the group members happened informally, face to face during collaborative activities (Principles 12, 13). The group member handling the projection took notes based on his/her own evaluations or instructions from other group members indicating what should be noted down or shown on the shared screen. This was problematic at times, as the more active students continuously expressed their thoughts and requests for the projection handler, while the less talkative participants would passively observe. Further, group members used digital tools to access online resources, shared documents and their Facebook group page. These tools enabled the communication between group members, in between the face-to-face sessions.

Primary flows of information between the group member (GM), group member handling the projection (GMp), tutor and the tools within the artifact ecology are shown in Fig. 4 below.



**Fig. 4.** Flow of information and communication channels

The primary mechanism used for information movement around the cognitive system is face-to-face interactions. Group members gathered around the downward pointing projection, brainstormed, identified learning issues, researched and acted as a united information decision hub (Principle 10). The high proximity between group members ensured that during collaborative learning activities information propagated in diversified forms. Physical artifacts, such as field notes or prototypes, or digital tools, such as smartphones and tablets, moved across group members facilitating information movement and discussion (Principle 8). A typical conversation between a group member and the handler of the projection included suggestions or instructions for action. This communication channel could also include the projection handler requesting clarifications concerning previous instructions from group members as shown in the example below (see Table 2 - “Channel 1”) (Fig. 4).

- GM 1: “Move to another concept, other factors are in the way of using buses and we can say what regular people can do to help with this!” [Reading from notes]
- GM 2: Oh! Write this down. [Turning head to the projection handler with instructions]
- GM 3: And it can concern all the types of bus users, for example older people.
- GM 1: Yes, and we can... Em...
- GM 4: Basically, is raising awareness.
- GM 1: Yes, yes. We can take it as a public problem. [Nodding her head]
- GM 4: How to write it down though? [Handler requests clarifications]
- GM 1: With whatever comes first in your mind. Keywords, short description. Just the way you explained this to us earlier.

Another mechanism for the propagation of information was Facebook where groups posted their ideas, resources or snapshots from physical artifacts and discussed about the product design process. The social networking platform was used particularly to support communication between face-to-face sessions. Sharing on the social networking platform ensured that the content was available to the rest of the group anytime, anywhere, turning the Facebook into an information buffer (Principle 11).

In the case of information transformation mechanisms, learners turned the physical artifacts into digital artifacts (Principle 9). While working towards the solution of the design problem, group members took notes and created sketches as initial steps of their solution. Using digital gadgets, such as tablets or their own smartphones, they took snapshots and shared them, transforming them into digital artifacts.

The initial description developed based on field notes and focus groups indicated that all groups kept tasks and learning goals in checklists for every session. The video analysis verified and extended the description revealing that one particular group – in addition to the checklist – enabled triggering points to scaffold their collaborative activities (Principle 14). In particular, in every session, groups discussed and reflected on their findings and crossed checked the issues that were completed, indicating “COMPLETE” next to each issue. An unfinished issue would cause the group another cycle of research, reporting and reflection.

**Table 2.** Summary for information flow channels of the cognitive system

Channels	Summary
1. Group member to projection handler	A group member expresses a suggestion that requires a reaction from the group member handling the projection, e.g. “open the report”, “note that down”. The projection handler receives the information and takes action by: <ul style="list-style-type: none"> <li>• Taking notes about the idea</li> <li>• Requesting further information or</li> <li>• Researching the idea further through online sources</li> </ul>
2. Group member to group member	The group members discuss regularly the problem, the solution and the procedure to construct the solution. These conversations take place during face-to-face collaborative activities or through Facebook in-between sessions.
3. Group member to tutor	The dialogue between a group member and tutor might be initiated in two ways: <ul style="list-style-type: none"> <li>• A group member requests guidance towards a particular aspect of the problem. The tutor provides additional triggers or hints, allowing the group to direct their own learning and discover the answers to their questions.</li> <li>• The tutor provides new material for the group to review and embed in their solution. The group members review material individually and collaboratively reflect on the new information.</li> </ul>

### 4.3 Artifacts

The artifact model focuses on the thorough analysis of individual artifacts that are deemed important within the cognitive system. In this case, the downward projection, mobile devices and Facebook emerged as key mediating artifacts (Principle 15) to analyze further and interpret how they supported the collaboration and coordination during the collaborative activities. The downward pointing projection acted as a focal point for group activities, keeping the group concentrated on the task at hand, distributing awareness. Mobile devices such as tablets and smartphones, played a

significant role mediating the transformation of physical artifacts into digital. For example, iPod was particularly used by Group 1 as a recorder to keep track of the ideas that were being discussed around the table, as in the example below.

- GM 1: Where is the iPod? Oh yes in the box. [Reaching for the iPod in the box.]
- GM 2: I think we should record whatever we discuss around the table because we will not remember everything.
- GM 1: Yes. Because now you all have said too many things and I could not take notes for all these. Wait a minute. [Searching an application in the iPod]
- GM 3: Doesn't matter. These were just thoughts that came out randomly, no formal ideas. Just preparatory phase. [Waving her hands]
- GM 1: Where is the...recorder? [Searching recorder application in the iPod]
- GM 2: Come on, I will do it. [Getting the iPod from Student 1]

Furthermore, Facebook constituted the primary communication and coordination medium for in-class and online interactions. As a coordination tool, Facebook captured the storyline of the group work, keeping a record of shared resources and issues discussed (Principle 18) [27]. In terms of scaffolding, the social networking platform offered the ability to categorize posts in themes or associate a post to an individual (Principle 16). The to-do list created in the previous session was either shared on Facebook or as a list in a shared document. Next to each item on the to-do list there was an indication of the learner or pair responsible for completing the task, thus, reducing the cognitive load of the group members to recall the roles assigned to each of them.

## 5 Discussion and Implications

Our study sought to illustrate the utility of DC and DiCoT as a tool for modelling interactions and interdependencies during collaborative learning activities in an artifact ecology. The findings showcase how the principles of DiCoT organized in three models – physical layout, information flow and artifacts – support the understanding of the learner-learner and learner-artifact interactions evident in the collaborative learning environment. Furthermore, the study highlights the affordances of the ecology of artifacts that support collaboration and cooperation during collaborative learning activities. The following sections elaborate on these findings and discuss implications of the study in relation to the literature.

### 5.1 DiCoT as a Modeling Tool

As claimed in the literature, DC is a well suited conceptual framework when dealing with technology-rich collaborative environments. The DiCoT framework explicitly emerged from the need to develop a methodology for DC analysis, using three models of behavior – physical layout, information flow and artifacts [8]. Extending the literature, one question this study sought to address was whether DiCoT is particularly well suited for modelling interactions and interdependencies during collaborative learning activities in an artifact ecology.

Considering the physical layout model of DiCoT, results illustrate that information is externally represented in the physical surroundings. The analysis demonstrated how the technological set-up (i.e., the artifact ecology) impacts the access to artifacts and the propagation of information through the cognitive system. The artifact and information flow models highlighted the distinguished roles that technological artifacts such as the downward projection, mobile devices and social networking platform (in this case, Facebook) play in coordinating activities and facilitating reflection of the product design process. Findings of this study support previous literature that DC can provide a lens for understanding collaborative learning activities in technology-rich spaces. Furthermore, the study adds to the validity of DiCoT as a well-suited modelling tool and a methodological and analytical tool for understanding complex interactions amongst learners, tasks, and tools in collocated technology-rich, learning environments.

## 5.2 Affordances of Artifact Ecology and Design Implications

Furthermore, the study sought to reveal the affordances of the ecology of artifacts that support collaboration and coordination during collaborative learning activities. A major affordance of the physical rearrangement of the ecology of artifacts is the close proximity between team members and artifacts, increasing awareness and supporting distributed cognition within the group. In a study by Sharp and Robinson [18], large areas namely “The Wall”, assisted in the coordination of an eXtreme Programming team’s activities. Team members could identify the progress of the collaborative activities from a distance, wherever in the room. In the present study the proximity between groups and tools assisted on increasing awareness by observing and listening to the issues raised by other group members. In one case, the large vertical surface distributed the status of the group work, while in the other case, the downward pointing projection created the necessary proximity to distribute awareness. We can conclude that tools perform differently across tasks and users, creating numerous possibilities for designing artifact ecologies and widening the horizons of research for in situ studies in a variety of contexts.

Furthermore, considering the flow of information within the artifact ecology we can suggest a number of design directions for the design of learning environments. Due to the set up in communications channels, there is currently a lengthy procedure if a group member expresses an idea to be recorded or explored through the projection, resulting in delays in the collaborative activities. What’s more, one or more group members might dominate the control of the projection through direct control of the mouse/keyboard or via multiple requests to the projection handler (see Sect. 4.2). Last, participants perceived natural to interact with direct and touch input with the projection on the table, suggesting a shift on what is nowadays perceived as natural interaction and transitioning towards tactile and gestural interaction styles. Based on the above, one potential design implication would be to switch the vertical projection with a surface computing system. That is, tabletops could help overcome the lengthy procedure of controlling the shared screen, reduce dominant actions, and increase the level of naturalness in the artifact ecology. There is already substantial work on using tabletops to support collaborative learning [30, 31]. There is also evidence that tabletops can help address the above-mentioned issues, moving attention away from a single input device

such as a mouse or keyboard [32], promoting equal participation and shared control [33]. Therefore, the present work points attention to surface computing as a technology that can expand the artifact ecology to provide further support for collaboration and learning. Based on our findings concerning the use of the shared projection, we can provide a couple of directions for design elements in tabletop application for integration in this or similar contexts:

- Enable collaboration on shared artifacts that can be loaded into a group project.
- Support pooling of information from previous meetings.
- Enable discussion and recording of alternative solutions.
- Allow tracking of decisions.
- Enable continuity in learners' interactions across time and space (face-to-face and online).

Furthermore, findings from this study seem to suggest that the use of secondary displays and interactive screens is necessary for sharing awareness and providing behavioral trigger factors. In particular, participants used mobile devices, to demonstrate supporting material, physical or digital, as well as take personal notes (see Sects. 4.1 and 4.3). This observation is consistent with Huang, Mynatt and Trimble's [22] showing that multiple displays can advance the distribution of cognition in a complex work space. However, based on the specific use of the mobile devices in the present study, we can provide few directions for implications for the design of mobile applications for integration in this or similar contexts:

- Support note taking and checklists applications with the ability to share or keep private.
- Allow tagging for organization and searching of completed and pending tasks.
- Present notifications linked to the group project and its progress.

### 5.3 Limitations and Future Directions

One limitation of this work is the lack of a detailed analysis of the dynamic and constantly changing artifacts that were encountered in the environment. For instance, the secondary displays in the environment changed constantly based on users' actions rather than being static representations during collaborative activities. In future work, this limitation can be overpowered by extending our analysis to include a temporal dimension, taking into account how users, tasks and tools change over time.

## 6 Conclusion

In the introduction we discussed how the different technologies available now at our fingertips can provoke new challenges on designing artifact ecologies. We also indicated the potential benefits of employing DC to understand the strengths and weaknesses of an artifact ecology designed to support collaborative learning activities. Through an in-class exploration, this work illustrated the utility of DC and DiCoT as a tool for modelling

interactions and interdependencies during collaborative learning activities in an artifact ecology. We demonstrated that DC is a well suited framework for understanding collaborative learning activities in technology-rich spaces and that it can be applied practically through DiCoT. DiCoT enabled the consistent mapping of the complex interactions and interdependencies within the artifact ecology and allowed us to showcase learner-learner and learner-artifacts interactions evident in the workspace. We further highlighted the affordances of the ecology of artifacts that supported collaboration and cooperation during collaborative learning activities and presented potential areas of improvement linked to design directions for tools to be used in similar contexts.

## References

1. Weiser, M.: The computer for the 21st century. *Sci. Am.* **265**(3), 94–104 (1991)
2. Jung, H., Stolterman, E., Ryan, W., Thompson, T., Siegel, M.: Toward a framework for ecologies of artifacts: how are digital artifacts interconnected within a personal life? In: Proceedings of the 5th Nordic Conference on Human-Computer Interaction: Building Bridges, pp. 201–210 (2008)
3. Bødker, S., Klokmos, C.N.: Dynamics in artifact ecologies. In: Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design, pp. 448–457 (2012)
4. Loke, S.W., Ling, S.: Analyzing observable behaviours of device ecology workflows. In: ICEIS (4), pp. 78–83 (2004)
5. Salomon, G.: Effects with and of computers and the study of computer-based learning environments. In: De Corte, E., Linn, M.C., Mandl, H., Verschaffel, L. (eds.) Computer-Based Learning Environments and Problem Solving, pp. 249–263. Springer, Heidelberg (1992)
6. Halverson, C.A.: Activity theory and distributed cognition: Or what does CSCW need to Do with theories? *Comput. Support. Coop. Work* **11**, 243–267 (2002)
7. Rogers, Y.: Coordinating computer-mediated work. *Comput. Support. Coop. Work (CSCW)* **1**(4), 295–315 (1992)
8. Blandford, A., Furniss, D.: DiCoT: a methodology for applying distributed cognition to the design of teamworking systems. In: Gilroy, S.W., Harrison, M.D. (eds.) DSV-IS 2005. LNCS, vol. 3941, pp. 26–38. Springer, Heidelberg (2006)
9. Hutchins, E.: *Cognition in the Wild*. MIT Press, Cambridge (1995)
10. Hollan, J., Hutchins, E., Kirsch, D.: Distributed cognition: Towards a new foundation for human-computer interaction research. *ACM Trans. Comput.-Hum. Interact.* **7**(2), 174–196 (2000)
11. Hutchins, E., Klausen, T.: Distributed cognition in an airline cockpit. In: Engström, Y., Middleton, D. (eds.) *Cognition and Communication at Work*, pp. 15–34. Cambridge University Press, New York (1996)
12. Nobarany, S., Haraty, M., Fisher, B.: Facilitating the reuse process in distributed collaboration: a distributed cognition approach. In: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, pp. 1223–1232. ACM (2012)
13. Rogers, Y.: *HCI Theory: Classical, Modern, and Contemporary*. Synthesis Lectures on Human-Centered Informatics. Morgan and Claypool, San Francisco (2012)
14. Beyer, H., Holtzblatt, K.: *Contextual design: Defining customer-centered systems*. Morgan Kauffman, San Francisco (1998)

15. Furniss, D., Blandford, A.: DiCoT modeling: from analysis to design. In: Proceedings of CHI 2010 Workshop Bridging the Gap: Moving from Contextual Analysis to Design, pp. 10–15 (2010)
16. Rajkomar, A., Blandford, A.: Distributed cognition for evaluating healthcare technology. In: Proceedings of the 25th BCS Conference on Human-Computer Interaction, pp. 341–350. British Computer Society (2011)
17. Rajkomar, A., Blandford, A.: Understanding infusion administration in the ICU through Distributed Cognition. *J. Biomed. Inform.* **45**(3), 580–590 (2012)
18. Sharp, H., Robinson, H.: Collaboration and co-ordination in mature eXtreme programming teams. *Int. J. Hum Comput Stud.* **66**(7), 506–518 (2008)
19. Hilliges, O., Terrenghi, L., Boring, S., Kim, D., Richter, H., Butz, A.: Designing for collaborative creative problem solving. In: Proceedings of the 6th ACM SIGCHI Conference on Creativity and Cognition, pp. 137–146. ACM (2007)
20. Rick, J.: Towards a classroom ecology of devices: interfaces for collaborative scripts. In: Workshop Proceedings of 8th International Conference on Computer Supported Collaborative Learning (CSCL 2009): “Scripted vs. Free CS collaboration: alternatives and paths for adaptable and flexible CS scripted collaboration” (2009)
21. Bardill, A., Griffiths, W., Jones, S., Fields, B.: Design tribes and information spaces for creative conversations. In: The 12th International Conference on Engineering and Product Design Education (2010)
22. Huang, E.M., Mynatt, E.D., Trimble, J.P.: Displays in the wild: understanding the dynamics and evolution of a display ecology. In: Fishkin, K.P., Schiele, B., Nixon, P., Quigley, A. (eds.) PERVASIVE 2006. LNCS, vol. 3968, pp. 321–336. Springer, Heidelberg (2006)
23. Looi, C.K., Wong, L.H., Song, Y.: Discovering mobile computer supported collaborative learning. In: Berge, Z.L., Muilenburg, L.Y. (eds.) The International Handbook of Collaborative Learning. Routledge, New York (2012)
24. Martinez-Maldonado, R., Clayphan, A., Ackad, C., Kay, J.: Multi-touch technology in a higher-education classroom: lessons in-the-wild. In: Proceedings of the 26th Australian Computer-Human Interaction Conference on Designing Futures: the Future of Design, pp. 220–229. ACM (2014)
25. Koschmann, T., Stahl, G.: Learning issues in problem-based learning: Situating collaborative information. In: CSCW Workshop on Technologies for Collaborative Information Seeking. ACM (1998)
26. Hmelo-Silver, C.E.: Problem-based learning: What and how do students learn? *Educ. Psychol. Rev.* **16**(3), 235–266 (2004)
27. Ioannou, A., Vasiliou, C., Zaphiris, P., Arh, T., Klobučar, T., Pipan, M.: Creative multimodal learning environments and blended interaction for problem-based activity in HCI education. *TechTrends* **59**(2), 47–56 (2015)
28. Vasiliou, C., Ioannou, A., Zaphiris, P.: Technology enhanced PBL in HCI education: a case study. In: Kotzé, P., Marsden, G., Lindgaard, G., Wesson, J., Winckler, M. (eds.) INTERACT 2013, Part IV. LNCS, vol. 8120, pp. 643–650. Springer, Heidelberg (2013)
29. Savery, C., Hurter, C., Lesbordes, R., Cordeil, M., Graham, T.: When paper meets multi-touch: a study of multi-modal interactions in air traffic control. In: Kotzé, P., Marsden, G., Lindgaard, G., Wesson, J., Winckler, M. (eds.) INTERACT 2013, Part III. LNCS, vol. 8119, pp. 196–213. Springer, Heidelberg (2013)
30. Higgins, S.E., Mercier, E., Burd, E., Hatch, A.: Multi-touch tables and the relationship with collaborative classroom pedagogies: A synthetic review. *Int. J. Comput.-Support. Collaborative Learn.* **6**(4), 515–538 (2011)
31. Dillenbourg, P., Evans, M.: Interactive tabletops in education. *IJCSCL* **6**(4), 491–514 (2011)

32. Harris, A., Rick, J., Bonnett, V., Yuill, N., Fleck, R., Marshall, P., Rogers, Y.: Around the table: are multiple-touch surfaces better than single-touch for children's collaborative interactions? In: Proceedings of the 9th International Conference on Computer Supported Collaborative Learning, vol. 1, pp. 335–344 (2009)
33. Ioannou, A., Zaphiris, P., Loizides, F., Vasiliou, C.: Let'S Talk About Technology for Peace: A Systematic Assessment of Problem-Based Group Collaboration Around an Interactive Tabletop. *Interacting with Computers*, iwt061 (2013)

# Assessing a Collaborative Application for Comic Strips Composition

Eleonora Mencarini<sup>1,2()</sup>, Gianluca Schiavo<sup>2</sup>, Alessandro Cappelletti<sup>2</sup>, Oliviero Stock<sup>2</sup>, and Massimo Zancanaro<sup>2</sup>

<sup>1</sup> University of Trento, Trento, Italy

<sup>2</sup> FBK-irst, Trento, Italy

{mencarini,gschiavo,cappelle,stock,zancana}@fbk.eu

**Abstract.** In this paper we present the evaluation of an application for the collaborative composition of comics using a pre-defined set of images and sentences. This study is an intermediate step to guide the design of a tablet application for supporting collaborative storytelling between two authors from different cultures and speaking different languages. For this purpose, we assessed the effectiveness of a constrained-text approach for comic composition in which sentences are selected from a library rather than written by the authors. Our results show that the constrained-text approach provides a satisfying form of co-narration, stimulating the authors to stay on topic, while using the available narrative material. The findings of this study have implications for the future design of collaborative storytelling applications for multilingual and cross-cultural scenarios.

**Keywords:** Collaborative storytelling · Teenagers · Creativity support index · Multilingual communication

## 1 Introduction

Stories are powerful educational tools for discussing cultural differences and diversity [8]. In the process of narrating a story, authors express their identities and make narrative decisions, characterizing the possible world they create [6]. This aspect becomes particularly important when the narrators are teenagers, as they are in the process of developing their own identity [5].

As ultimate goal of our research, we aim at designing and developing a tool to help teenagers who do not share a common language to write stories collaboratively. The work presented in this paper is an intermediate step toward this goal and describes a tablet application for remote collaborative storytelling, in which ready-made language expressions are selected from a library, rather than freely written by the authors. The constrained-text interaction has been devised in order to offer an alternative to real-time automatic and human translation in multilingual scenarios. In order to explore how the constraints provided by our technology influence users' satisfaction and creativity, we designed an experimental study with teenagers (15–17 years old) who share the same language and asked them to create a story both in a

constrained-text and in a free-text version. The results suggest that composing a story with constrained text is a valuable alternative to other approaches to remote collaborative storytelling, such as free writing. In particular, our study shows that the constrained-text approach supports collaborative storytelling by providing authors with a limited yet satisfying level of expressiveness.

## 2 Collaborative Storytelling

Collaborative storytelling has long been an educational activity to support reflection on identity [4]. Several attempts have been made in designing technological devices to assist the narrative process. Social networking infrastructures such as Facebook [2], or ad-hoc mobile applications such as MAHI [5], have been used to support digital storytelling in remote, showing how teenagers represent themselves and negotiate their identities online. Some studies in HCI and Education have investigated how teenagers express their narrative ideas in the realization of digital contents, such as videogames, describing the educational benefits of digital storytelling through creative activities [6]. Finally, other studies have investigated the benefits of adopting collaborative and co-located storytelling for cross-cultural mediation. These latter studies have shown that collaborative storytelling can be used to facilitate conflict resolution and to support mutual understanding between authors with different cultures [8].

Our study enriches this last research line, investigating remote collaborative storytelling through the use of constrained-text with the eventual goal of fostering multilingual communication and cross-cultural dialogue.

## 3 Design Objectives and System Description

The 3 main design objectives of this study were: (i) to ensure the creation of meaningful stories; (ii) to support remote collaboration between two authors and (iii) to design a system that in the future can allow remote collaboration also between authors who do not share the same language.

Regarding the first objective, we provided content and images for the development of at least 4 different stories on the topic of ecology. To manage remote collaboration, we decided to adopt a turn-based approach, where the authors alternate in composing the story by intentionally giving the turn one to the other. Regarding the multilingual support, although the use of a common language (such as English as *lingua franca*) has itself an educational value, storytelling in a non-native language can be difficult and it may hinder collaboration if the two authors do not share the same level of knowledge of the language. Only high proficiency in the language on both sides could ensure full expressiveness and richness in the conversation. Hence, two other main approaches might be used to tackle the challenge of multilingual communication: a free-text approach, where the author freely enters text in her own language, to be translated by a person or by an automatic translation system; or a constrained-text approach, where a number of sentences, previously translated and paired, are made available to the authors for selection. The problems with the former option are that the mediation of a human translator might influence the outcome of

the storytelling (as in [8]), and that the current tools for automatic translation are still far from being usable on more than a single word or sentence, and their use might impact the collaboration experience [3]. While the problem of the latter option is that the limited content may hinder authors' expressiveness. However a constrained-text approach has the advantage to minimize the translation process, since the pre-defined sentences can be previously paired in two (or more) languages. The present study is specifically focused on understanding and assessing the impact of a constrained-text approach on remote collaborative composition of comics using a tablet application.

In designing the application, we opted for comic strips as well-known narrative genre among teenagers. The proposed technology is a tablet application that provides a library of images (42 in total) for backgrounds, characters and objects, as well as a repository of predefined language expressions (151 in total) for greetings (e.g. "Hello!" "Bye-bye!"), statements (e.g. "Look at that poor animal, it has been caught by something!"), questions, etc. that the users can arrange to compose the comic strips (Fig. 1). The number of items (both graphical and lexical) allowed a manageable navigation of the interface, requiring a few horizontal scrolls.

The story was collaboratively created by two users remotely located, each one using her tablet. The authors worked in turns in composing the story, while the interface was automatically synchronized between the two tablets.



**Fig. 1.** A screenshot of the application in the constrained-text condition, showing a story partially created (upper part of the figure) and the library of language expressions (lower part).

In order to investigate the limitations that this approach may bring to storytelling, we developed a slightly different version of the application, which allows free-text input for dialogues and captions. In all other respects (such as the graphics) the two versions share the same features.

## 4 The Study

### 4.1 Objective

A study was designed to compare the constrained-text approach with a traditional free-text condition in a monolingual setting. The objective was to understand whether the

constrained approach allowed users to construct meaningful stories (despite the given topic and the limited resources) and whether it ensured a satisfactory experience.

Our hypothesis was that the constrained-text approach could provide a satisfactory support to remote collaborative storytelling with respect to the free-text condition. We consider the constrained-text version as a satisfactory support to storytelling if it guarantees a reasonable level of expressiveness and allows the creation of meaningful stories.

## 4.2 Measures

During the study both objective and subjective measures were collected and analyzed. Objectives measures consisted of the logs of the activities on the interface, which pertained the interactions with the tablet as well as the story creation. Measures about the interaction regarded the number of turns, their duration, and the number of operations (addition and removal of elements) performed in each turn. Information about the story creation included the number of scenes and of elements used.

Subjective measures included a 3-item questionnaire investigating users' satisfaction with the story, with the collaboration on the task and the perceived difficulty of using the application (Table 2). Each item was scored on a 5-point Likert Scale. Furthermore, we adopted the Creativity Support Index (CSI) to measure the perceived usefulness of the application in supporting the creation of stories. The CSI is a psychometric survey designed to investigate the ability of a digital tool to support a creative process [1]. CSI is composed of 12 items and 6 different scales for measuring the following dimensions of creativity: Collaboration, Enjoyment, Exploration, Expressiveness, Immersion and Result Worth Effort.

## 4.3 Procedure

The study was designed as a within-subject experiment with the constrained-text (CT) application as main condition, and the free-text (FT) version, which allows free text in dialogues and captions, as control condition.

The study was conducted in a 3rd year high school class, with 20 participants (11 females and 9 males; mean age = 15.85 years), all speaking the same language (Italian). The activity, although encouraged by the literature teacher, was not considered part of the normal class activity and participants did not receive a grade for their performance.

Before starting the storytelling activity, the participants were told about the goal of the task (i.e. "create a story on the theme of sustainability") and were trained in using both versions of the application. During the study, each participant received a tablet and was anonymously paired with another student (for a total of 10 pairs). All pairs worked on the creation of a story using the application both in the CT and FT condition, and the starting condition was randomly counterbalanced over the pairs. In each condition, the same incipit of the story was given in advance: a scene with two boys in a schoolyard and a caption saying: "Bill and John meet after school..." For each condition, participants were given 12 min to conclude the story and afterwards they were asked to fill out the questionnaires.

## 5 Results

### 5.1 Quantitative Data Analysis

**Structure of the Stories and Interaction Between the Authors.** Each pairs used on average 9 turns for completing the story in the CT condition vs. 7 turns in the FT condition. Each turn lasted on average 2 min, but standard deviation was higher for the CT condition, indicating a more variable time usage. This might be explained by the fact that it is quicker to compose a scene when the graphical and constrained-text elements are identified, while it might take longer when the users spend time navigating through the different categories. On average, the stories were composed of 4 frames and the number of graphical elements used was similar in both conditions. When using the constrained-text application, the participants tended to use a higher number of textual boxes. This might be explained by the fact that in the free-text condition the participants could use a single text box for expressing more concepts. However, as shown in Table 1, metrics from the system logs were not statistically different between conditions (Wilcoxon Signed Ranks Test,  $p > .05$ ).

**Table 1.** Interaction and story structure measures registered by the system logs – Mean (SD)

	Constrained-text condition (CT)	Free-text condition (FT)
<b>Stories</b>		
# of frames	4.1 (1.6)	3.8 (1.2)
# of graphical elements	10.1 (2.8)	10.6 (5.1)
# of textual elements	14.5 (6.8)	10.7 (2.9)
<b>Interaction</b>		
# turns	9.4 (4.97)	7.4 (4.35)
Turn duration (sec)	143 (250)	132 (81)
# operations per turn	14 (10)	13 (10)

**Satisfaction and Difficulty.** From the questionnaire investigating the users' impressions of the system, we could see that participants were generally satisfied with the story created and with the collaboration, and no statistical differences were observed between the two conditions (Table 2). Yet, the constrained-text application was considered more difficult to use, although both scores are below the midpoint of the Likert scale corresponding to "neutral".

**Creativity Support Index Analysis.** Table 3 shows the average CSI scores, including the average factor counts that express which factors are considered more important for the creativity activity regardless to the specific technology used (highest value is 5), and the average factor score, representing how well the specific version of the application

**Table 2.** Responses to the satisfaction and perceived difficulty questions on a 5-point Likert scale from 1 (Not at all) to 5 (Very much).

Items	CT	FT	F and p value
I was satisfied with the story	3.4 (0.9)	3.5 (1.1)	$F_{1,19} = 0.17; p > .05$
I was satisfied with the collaboration	3.95 (0.9)	3.7 (1.1)	$F_{1,19} = 1.34; p > .05$
I found the application difficult to use	2.75 (1.2)	1.2 (0.4)	$F_{1,19} = 11.47; p < .01$

supports these factors (highest value is 20). The scores indicate that *Collaboration* is by far the most important factor considered by the users with respect to collaborative remote comic writing, while *Results Worth Effort* and *Enjoyment* were regarded as the less relevant to the task.

The overall CSI score was calculated by multiplying each factor scores by the related factor counts. It represents how well the tool supports the factors that are considered as most important to the storytelling task. The CSI score for the constrained-text application was 65.1 while the free-text application received a score of 75.57. These scores suggest a reasonably good creativity support for both applications [1]. A repeated-measure ANOVA shows that they are statistically different ( $p < .05$ ), with the higher score for the FT condition (Table 3).

**Table 3.** CSI Results for CT and FT conditions – Mean (SD). Resulting p values are adjusted for multiple comparisons with Bonferroni correction.

Scale	Avg factor counts	Avg factor score		F values	p values
		Constrained-text	Free-text		
<i>Collaboration</i>	4.45 (0.83)	14.64 (4.49)	15.71 (3.85)	$F_{1,19} = 0.87$	$p > .05$
<i>Immersion</i>	2.60 (1.14)	10.86 (4.91)	12.79 (3.72)	$F_{1,19} = 7.65$	$p < .05$
<i>Exploration</i>	2.20 (1.47)	11.43 (5.54)	14.07 (5.19)	$F_{1,19} = 10.61$	$p < .01$
<i>Expressiveness</i>	2.15 (1.39)	10.21 (5.27)	15.36 (4.84)	$F_{1,19} = 24.52$	$p < .01$
<i>Enjoyment</i>	2.05 (1.70)	14.36 (4.94)	16.50 (3.26)	$F_{1,19} = 5.66$	$p < .05$
<i>Results worth effort</i>	1.55 (1.57)	13.00 (4.97)	14.79 (4.78)	$F_{1,19} = 4.57$	$p < .05$
<b>Overall CSI score</b>		<b>65.1 (22.76)</b>	<b>75.57 (17.39)</b>	$F_{1,19} = 7.73$	$p < .05$

Looking at the individual factors and how they were rated across the two conditions, we can observe that both versions received high values for the *Collaboration* dimension. A MANOVA indicates a difference between conditions (Pillai's trace,  $V = 0.61$ ,  $F_6, 14 = 3.7$ ,  $p < .05$ ); specifically, all the scores except for *Collaboration* are higher for the FT condition (Table 3). This suggests that the free-text application provides more support as creative tool in the majority of the factors investigated but *Collaboration*. This result is relevant considering that collaboration was rated as the most important factor in collaborative comic writing and it was one of the main design objectives.

## 5.2 Qualitative Analysis of the Stories

The content of the 20 stories was assessed with a qualitative approach. We used the dimensions of (i) *relevance to the topic* (i.e. ecology), (ii) *narrative tension* as the presence of a problem and its resolution [7] to mark the meaningfulness of the stories, as well as (iii) *level of interactivity* between the two characters [6] to assess the level of expressiveness of each story.

**Relevance to the Topic.** All the stories in the CT condition are centered on the topic of ecology as required by the task. This is somehow expected since they were built out from a pre-defined library of language expressions on this topic. Nevertheless, it was possible, though not simple, to build off-topic stories with the material provided. On the contrary, several stories in the FT are off-topic (6 out of 10). In 2 of these cases, the FT condition was performed after the CT, so by that time the authors had already created an on-topic story.

**Narrative Tension.** Narrative tension is only hinted in some stories and absent in many others. In CT condition, 4 stories have a goal or a solution (such as freeing a squirrel caught in a net) and 5 in FT condition. Still, among the CT stories without a goal or solution, 3 have a clear moral (such as, “it is because everybody behaves like you that we are in trouble”).

**Level of Interactivity.** Stories in both conditions contain several dialogues between the two characters and textual elements were widely used in both conditions. However, the general impression is that the plots are simple. That might be a consequence of the limited amount of time available for the tasks.

## 5.3 Discussion

**Satisfaction and Gratification.** Supporting our hypothesis, the subjective data from the questionnaires show that the constrained-text application provided a satisfactory support to storytelling and to collaboration at a level similar to the free-text version. Unsurprisingly, the participants favored the FT application, which they found to provide better creativity support for authors engaged in storytelling. It is worth noting that the FT version was preferred along all dimensions but collaboration, and that collaboration was indeed scored as the most important aspect of the task.

**Content of the Stories (relevance to the given theme).** Although the low quality of the stories does not allow a full analysis of the differences in the two conditions, a noticeable difference was that the constrained-text approach seems to encourage on-topic stories and this may be considered as a positive aspect. Moreover, the objective measures (number of scenes and objects used) did not differ significantly, suggesting that the structural level of the stories is similar in the two conditions.

**Collaboration.** The analysis of the objective measures suggests that the two applications had similar complexity and overall the turn-exchanging mechanism was easy to use. Yet, in the FT condition, we observed that participants often enacted an

appropriation practice: they used the free text to discuss and find an agreement on the story, when they reached an agreement they would delete the text that therefore do not appear in the final version of the stories. Although this practice does not seem to have improved the quality of the stories, the possibility to communicate seems to be needed by the authors. As future work, we plan to add such functionality in the constrained-text application providing a similar constraint-based mechanism.

## 6 Conclusions

This paper is an intermediate step toward the design of an application for supporting remote collaborative storytelling between people from different cultures and speaking different languages. In the study presented in this paper, we compared a constrained-text version of an application for composing comic stories with a free-text version of the same application. Although the free-text application provided a higher level of expressiveness and ease of use, the constrained-text approach allowed the creation of simple, yet meaningful, stories. Therefore, the study presented in this paper provides the basis for future work to investigate the use of the constrained-text approach for supporting collaborative storytelling in multilingual setting.

## References

1. Cherry, E., Latulipe, C.: Quantifying the creativity support of digital tools through the creativity support index. In: TOCHI 2014, vol. 21, no. 4, pp. 1–25. ACM (2014)
2. Davies, J.: Facework on Facebook as a new literacy practice. Comput. Educ. **59**(1), 19–29 (2012). Elsevier, Amsterdam
3. Gao, G., Xu, B., Cosley, D., Fussell, S.R.: How beliefs about the presence of machine translation impact multilingual collaborations. In: Proceedings of CSCW 2014, pp. 1549–1560. ACM, New York (2014)
4. Luwisch, F.E.: Understanding what goes on in the heart and the mind: learning about diversity and co-existence through storytelling. Teach. Teach. Educ. **17**, 133–146 (2001). Elsevier, Amsterdam
5. Mamkina, L., Miller, A.D., Mynatt, E.D., Greenblatt, D.: Constructing identities through storytelling in diabetes management. In: Proceedings of CHI 2010, pp. 1203–1212. ACM, New York (2010)
6. Robertson, Judy, Good, Judith: Supporting the development of interactive storytelling skills in teenagers. In: Pan, Zhigeng, Aylett, Ruth S., Diener, Holger, Jin, Xiaogang, Göbel, Stefan, Li, Li (eds.) Edutainment 2006. LNCS, vol. 3942, pp. 348–357. Springer, Heidelberg (2006)
7. Theune, M., Linssen, J., Alofs, T.: Acting, playing, or talking about the story: an annotation scheme for communication during interactive digital storytelling. In: Koenitz, H., Sezen, T.I., Ferri, G., Haahr, M., Sezen, D., Çatak, G. (eds.) ICIDS 2013. LNCS, vol. 8230, pp. 132–143. Springer, Heidelberg (2013)
8. Zancanaro, M., Stock, O., Eisikovits, Z., Koren, C., Weiss, P.L.: Co-narrating a conflict: an interactive tabletop to facilitate attitudinal shifts. In: TOCHI, vol. 19, no. 3, 24, pp. 1–30. ACM, New York (2012)

# Augmenting Collaborative MOOC Video Viewing with Synchronized Textbook

Nan Li<sup>(✉)</sup>, Łukasz Kidziński, and Pierre Dillenbourg

CHILI, EPFL, Lausanne, Switzerland

{nan.li,lukasz.kidzinski,pierre.dillenbourg}@epfl.ch

**Abstract.** We designed BOOC, an application that synchronizes textbook content with MOOC (Massive Open Online Courses) videos. The application leverages a tablet display split into two views to present lecture videos and textbook content simultaneously. The display of the book serves as peripheral contextual help for video viewing activities. A five-week user study with 6 groups of MOOC students in a blended on-campus course was conducted. Our study in this paper reports how textbooks are used in authentic MOOC study groups and further explores the effects of the book-mapping feature of the BOOC player in enhancing the collaborative MOOC learning experiences.

**Keywords:** MOOC · Peripheral display · Contextual help · Collaborative learning

## 1 Introduction

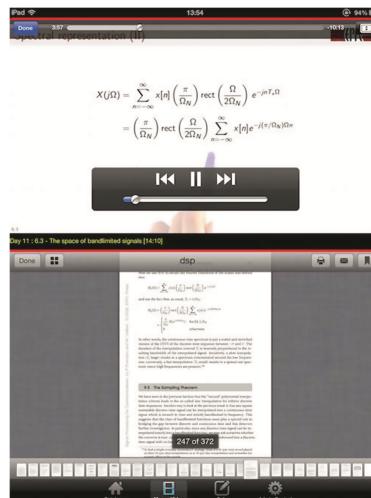
Massive Open Online Courses (MOOCs) are in recent years growing in popularity. Popular platforms such as Coursera and edX typically replicate traditional classroom pedagogy online, featuring with video lectures, quizzes, tutorials, discussion forums and wikis. Among these learning components, lecture videos play a central role in MOOC learning; the forums provide peer-to-peer support analogous to that in the classroom, and other components mainly serve as supporting resources. Note that textbooks are not essential for MOOCs. Their role as references is perhaps partially displaced by the lecture videos. This does not mean textbooks are not useful in MOOC learning. In a MOOC taught with a companion textbook, Belanger et al. [1] found that the students often spontaneously identified related content in the textbook, then shared and discussed them in the forum. This finding does not only exhibit that textbooks still function as potentially effective supporting materials for MOOCs, but also indicates the potential needs for providing the students with supervised book-to-video references.

Apart from textbooks, study groups, which is a common collaborative learning setting at schools and universities, are also suppressed by MOOCs, particularly by its massive and distributed nature. Compared to online forums and online groups,

collocated study groups allow more intimate discussions. In our recent work [4], we have shown that students are enthusiastic about studying MOOC in face-to-face groups. We also explored how students studied in different group arrangements, i.e. watching videos on individual displays or on a shared display. The key question to be addressed by this paper is, however, about a combinational usage of study group and textbook. We are interested in how the students use textbooks in their groups. Particularly, when referenced textbook content is linked to the videos, how does it affect the students' collaborative learning behaviors?

Textbooks are known to offer more comprehensive learning materials. When the books are linked to videos, the support becomes contextual help. Such help has been recognized as effective means for learning graphical interfaces [2]. In fact, most of the contextual help literature dealt with step-by-step web tutorials, and only several investigated video-based tutorials. Pause-and-Play [5] proposed a method to detect task-performing events in the video and link them with user actions in the target application as the user tried to imitate the procedure. This method avoids manually switching between the user context and the online tutorial. ToolClips [7] embedded video tutorials as contextual assistance for tool functionality understanding in a software application; FollowUs [6] demonstrated that video tutorials could be enhanced by multiple demonstrations of tutorial videos from other community members. These projects endeavored to create links between graphical interface to be learned and video tutorials. Practicing the software is the main activity that renders the context whereas the videos are seen as external helps. However, situations are poles apart in the MOOC scenario, where video viewing is the central activity. Additionally, video viewing is more passive and engaging, capturing most of the learners' attention.

In this paper, we evaluate BOOC, a MOOC video player that synchronizes lecture videos with relevant textbook pages in an authentic study group setting. The book pages are displayed in the periphery to minimize distractions. The paper contributes



**Fig. 1.** The BOOC player (Color figure online)

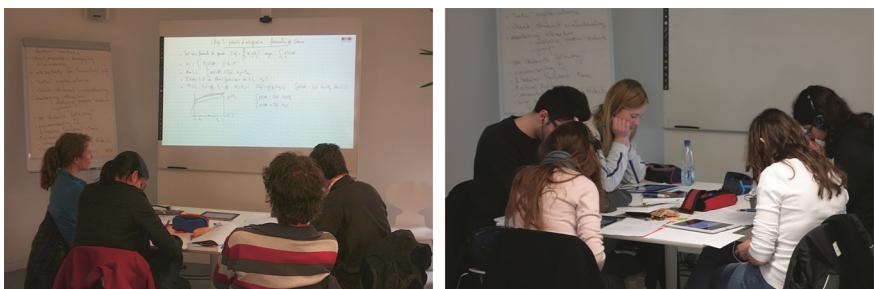
in understanding the roles of textbooks, especially when used as peripheral contextual help in MOOC study groups. We reveal how the students leverage the usage of books and peer discussion to study MOOC in groups.

## 2 The BOOC User Interface

As shown in Fig. 1, the BOOC user interface is separated into two parts, a video controller plays the lecture video and a PDF controller manages page navigations. Each controller has a horizontal status indicator, both of which are initially red, indicating the contents are synchronized. Double clicking on the PDF controller toggles the synchronization. Desynchronized PDF controller will have green indicator, which contrasts the other one in red, suggesting asynchronicity.

The two controllers are synchronized by default. If the video content is beyond the textbook, the PDF controller is greyed out, indicating that no textbook pages are relevant at the moment. In case multiple pages are related to the same video segment, the most relevant one is presented. Others have only their page numbers shown as yellow text in the middle of the screen. Students can navigate to those pages at their own effort, if they wish.

Videos and books are dually mapped. Users can also navigate through the digital book to get the corresponding video explanations, if available. In case of multiple mappings, the system pops out a list of other relevant videos for selection.



**Fig. 2.** CC (left) and DD (right) group conditions

## 3 User Study

Our study was based on an undergraduate-level MOOC offered by our university at Coursera, *Numeric Analysis*. The on-campus version of the course was arranged as 7-week flipped teaching with MOOC plus 8-week traditional teaching. In the flipped teaching period, the students were required to watch videos and solve quizzes at home. Classroom sessions were reserved for exercises and advanced tutorials. We recruited 25 volunteered on-campus students (8 females/17 males) from the course. Each subject was compensated 150 SFr. together with a print companion textbook for participating 5 weekly study sessions. Since it is common for students to study with familiar fellow students, we allow them to spontaneously form 6 groups by themselves.

Each group met weekly around a table to study together. Activities typically included watching videos and solving quizzes. Half of the groups shared a tablet that was connected to a beamer for video viewing, and we call this condition CC (Centralized display Centralized control, see Fig. 2 left). Students from the rest groups watched videos on her own tablet (borrowed from us) independently with earphones, and this condition is referred to as DD (Distributed display Distributed control, see Fig. 2 right). All the groups were composed of 4 students, except one DD group has 5 participants. We believe the two group conditions both naturally exist in real life, so we would like to understand how the groups study and use books accordingly. In other words, in spite of the group conditions, the study is more of an exploration rather than a comparison of experimental designs.

In addition, we did not instruct the students how to watch videos or collaborate. They were encouraged to behave naturally. However, we manipulated the forms of the video player. The subjects watched the videos for the first three weeks with a normal video player (without PDF); the BOOC player with book mappings was used only for the last two weeks. To our knowledge, there was no reliable automatic way to map book content with videos precisely, so we made it manually with the help of a designated teaching assistant of the course. Each week's videos (less than 10) took around one hour to finish the mapping.

For the data collection, we have video interaction logs obtained from the tablet video player. We also videotaped the study sessions so as to analyze the book usage and speech from the recordings. Two cameras were employed to capture both the front and rear view of the group interactions. In addition, there were weekly recurring post-experiment questionnaires with repeated questions regarding group collaboration, quality of discussion as well as book usage. Each session was followed by a semi-structured interview so as to obtain deeper insights of their learning experiences.

## 4 Results

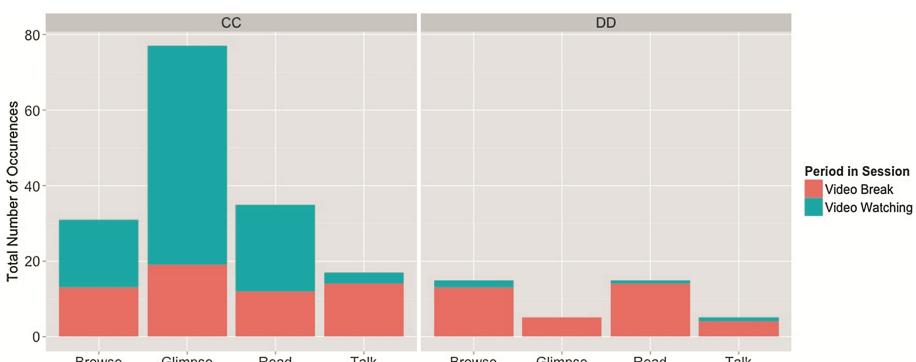
The event logs recorded the interactions on the BOOC player, but print book interactions as well as collective interactions involving multiple subjects need to be manually coded from the video recordings. One of the authors coded 440 book interactions found from video recordings of 30 study sessions (2–3 h each), and identified the following interactions modes:

- **Browse.** *Turning the print book rapidly to look for relevant pages.*
- **Glimpse.** *Glancing at the book to follow the video (mostly) or quiz. Such interactions typically lasted shortly, and the students' eyes usually quickly moved back-and-forth between the book and the video/quiz sheets.*
- **Read.** *Resting the eyes on the print book page for longer time to read text.*
- **Turn.** *Turning pages on the PDF book.*
- **Scroll.** *Scrolling the PDF book to view different parts of the page.*
- **Zoom.** *Zooming in/out to see details in the PDF book.*
- **Talk.** *Talking to other members with the book, sometimes with deictic gestures.*

Among these modes, *Browse*, *Glimpse* and *Read* are individual interactions on the print book; *Turn*, *Scroll* and *Zoom* are associated with the embedded PDF in the BOOC player. An interaction becomes collective if it is performed with conscious awareness of multiple group members, e.g. when they have shared interest in the book or discuss the book. *Talk* is a collective verbal interaction by definition. The presented interaction modes can be easily identified from the videos with the help of both front view and rear view recordings. In principle *Glimpse* and *Read* also apply to the PDF, but it is difficult to tell from the video recordings without eye tracking, since the videos and PDF reside in parallel.

#### 4.1 Textbook Usage in Pre-BOOC Sessions

In the first three weeks, the students watched lecture videos on normal tablet video players, and their companion textbooks were all placed on the table. Each week, a new sequence of videos was available, and we found that the students usually watched them in order. As soon as a group finished watching a video, they often had short discussions about the just-watched video or the associated quizzes before starting the next video. This period of time is referred to as *video break*, as compared to *video pause*, which refers to the time when a video is paused during its watching session. The DD students tended to synchronously start the videos and also tried to finish the videos at the same time so as to start discussions together. More thorough discussions about this phenomenon are presented in [4]. The book was not essential for studying the course, so the usage was voluntary. Not every group had used books in every session, and our goal was not to predict the usage. Rather, our interest lies in the interaction modes when the book was used. Figure 3 illustrates the frequency of each print book interaction mode of the DD and CC groups in the first three weeks. We find that the time period when the book interactions occurred differed significantly between CC and DD ( $\chi^2(200,1) = 34.98$ ,  $p < .0001$ ,  $\phi = 0.43$ ). The CC groups had more balanced usage of book between during the video break and video watching, while the DD students mostly used the textbooks in video breaks. This was especially notable for the *Glimpse* action, which indicated the student was following the videos with textbooks in parallel. The frequency is visibly high in the CC but none in the DD. A possible



**Fig. 3.** Frequencies of the print book interaction modes in pre-BOOC sessions

explanation could be the DD students did not want to break their video watching synchronicity for using the book.

Overall there were 20 occasions where the students grabbed the book on the table and looked for relevant content while watching videos (*Browse*). 32.5 s on average were spent on each *Browse*. In 3 out of the 20 *Browse* occasions, the students failed to find the intended information on the book. Collective book interactions were only in the form of *Talk* with print books. Usually we found a student asked questions to others with the book or read aloud the book to the group. *Talk* interactions account for respectively 10.6 % and 12.5 % in CC and DD and mostly (82.3 % and 80.0 %) happened during video break.

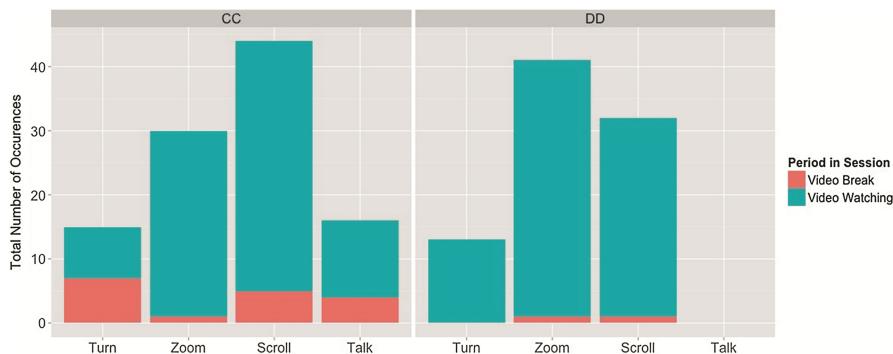
During the semi-structured interviews, we tried to understand why some of the students did not use the book during the interviews. Main reasons include (1) they were afraid of loosing time in looking for information in the book (2) they did not know exactly what is not clear while watching videos (3) the lectures were easy, so the videos are sufficient for learning (4) they prefer to ask the groups, which usually solved their problems. Feedbacks (3–4) reflect more of contextual constraints in the collaborative MOOC learning setup, where the students' doubts about video content depend on the course itself. They may prefer to address the problems through discussions in the group rather than the book. Feedbacks (1–2) as well as the observations presented before confirm the potential needs for peripherally displaying book references, so that the students can quick judge the usefulness of book content.

## 4.2 Textbook Usage in BOOC Sessions

We deployed the BOOC player with book mappings in the last two weeks of the study, but students were still asked to take out their print textbooks during the study sessions for potential usage. In fact, only one student had used the print book in the last two weeks, because he favored the tangibility of the print book. Still, he used the PDF mapping in BOOC to help quickly navigate to the pages.

To analyze the usage of the peripherally displayed synchronized PDF, we first summarize the usage scenarios collected from the questionnaire as below: (1) **Extended Knowledge**: when they saw another explanation of a concept with detailed theorems and examples (2) **Alternative Presentation**: when the teacher was talking too fast or the videos were not visibly clear (3) **Information Confirmation**: when the students had doubts about certain concepts and need to confirm their understanding (4) **External Help**: when none of the group members knew the answer or when they were arguing about certain concepts. Among these usage scenarios, the advantage of the peripheral design is especially notable in (1), where the students had no explicit needs of help – They actually discovered helpful information in the peripheral display of the book with serendipity.

Figure 4 depicts the frequency of book interactions by mode with BOOC interface. It is not surprising to see that in both conditions, the interactions predominantly happened during video watching rather than during video break due to the book-mapping feature. What is more interesting is that the DD students used the book frequently during video watching, which seldom happened before. A possible explanation could be the



**Fig. 4.** Frequencies of PDF interaction modes in BOOC sessions

synchronized PDF increased the visibility and accessibility of the potentially useful information in the book. As a result, the students were offered better opportunities to address their situational needs without the fear of loosing synchronicity.

For the students in the CC condition, the most notable change after the introduction of BOOC is the increased frequency of collective activities, especially during video watching. The proportion of collective activities doubles to 20.9 %. If we count only *Talk* interaction, the proportion is 15.2 %, with 75 % happened during video watching, as compared to 17.7 % in the pre-BOOC sessions. The reason behind the increments, we believe, is that the shared display of synchronized book content increased shared attention, so that the students could have more chances to collaborate with the book. One may wonder why the increment of book discussion matters. In fact, when the quality of discussion (5-point Likert scale ratings obtained from weekly questionnaires) in the CC groups was predicted, it was found that the frequency of *Talk* interactions during video watching ( $\beta = 0.24$ ,  $p < .05$ ), the proportion of speech during video break ( $\beta = 5.58$ ,  $p < .001$ ) were significant predictors. The statistics were computed from a mixed-effect multiple-regression model where the students nested in groups were modeled as random effects. The overall model fit in terms of R-squared was 0.48, and the marginal R-squared contributed by the fixed effects was 0.21. Significance could also be obtained if we fitted in the model with the frequency of all collective interactions or the duration of *Talk* interaction, instead of *Talk* frequency. This finding indicates that more discussions with book or collective book usage in general may increase discussion quality.

Remember that when we designed the BOOC interface, we enabled dual mapping between the PDF and the video. During the experiment, linking from book pages to videos were never intended, since watching videos was their main activity, not reading books. The students were sometimes annoyed due to abrupt video changes when they accidentally swiped the PDF to a page with a different video mapping. We argue dual mapping might be useful when used at home than in time-bounded group study sessions, where the students' activities are much centered on watching videos.

## 5 Conclusion

In this paper, we presented how textbooks were used in two common formats of authentic MOOC study groups. Particularly, we summarized the students' feedback on why print textbooks were not widely accepted and when the BOOC interface, which was essentially a peripheral contextual help display, was considered useful in the group study sessions. We also delivered an empirical categorization of book interaction modes during study group sessions. These qualitative findings increased our understanding of the role of the textbooks in collaborative MOOC learning setup.

In addition, we discussed how the book usage was changed since the introduction of BOOC player in a quantitative way. Most importantly, we found the BOOC player increased collective book usage for the CC groups, which in turn enhanced the students' discussion quality significantly. These findings further revealed the effectiveness of the peripheral book display as it increased mutual awareness of contextual information. Even in the DD condition, the display worked in the sense that it provided situational help to the students without losing much synchronicity, which is essential in the DD style study groups [4]. Future work includes replicating similar interfaces to the more dynamic online platforms, and contextual helps other than textbook may be employed to facilitate students' learning as well.

## References

1. Belanger, Y., Thornton, J., Barr, R.C.: Bioelectricity: A Quantitative Approach–Duke University's First MOOC. Duke Univeristy, Durham (2013)
2. Chilana, P.K., Ko, A.J., Wobbrock, J.O.: LemonAid: selection-based crowdsourced contextual help for web applications. In: Proceedings of CHI 2012. ACM (2012)
3. Chi, P.Y., Ahn, S., Ren, A., Dontcheva, M., Li, W., Hartmann, B.: MixT: automatic generation of step-by-step mixed media tutorials. In: Proceedings of UIST 2012. ACM (2012)
4. Li, N., Verma, H., Skevi, A., Zufferey, G., Blom, J., Dillenbourg, P.: Watching MOOCs together: investigating co-located MOOC study groups. *Distance Educ.* **35**, 217–233 (2014)
5. Pongnumkul, S., Dontcheva, M., Li, W., Wang, J., Bourdev, L., Avidan, S., Cohen, M.F.: Pause-and-play: automatically linking screencast video tutorials with applications. In: Proceedings of UIST 2011. ACM (2011)
6. Lafreniere, B., Grossman, T., Fitzmaurice, G.: Community enhanced tutorials: improving tutorials with multiple demonstrations. In: Proceedings of the CHI 2013. ACM (2013)
7. Grossman, T., Fitzmaurice, G.: ToolClips: an investigation of contextual video assistance for functionality understanding. In: Proceedings of CHI 2010. ACM (2010)

# EXCITE: Exploring Collaborative Interaction in Tracked Environments

Nicolai Marquardt<sup>1()</sup>, Frederico Schardong<sup>2</sup>, and Anthony Tang<sup>2</sup>

<sup>1</sup> University College London, Gower Street, London, UK

[nicolai.marquardt@acm.org](mailto:nicolai.marquardt@acm.org)

<sup>2</sup> University of Calgary, 2500 University Drive, Calgary, NW, Canada

[frede.sch@gmail.com](mailto:frede.sch@gmail.com), [tonyt@ucalgary.ca](mailto:tonyt@ucalgary.ca)

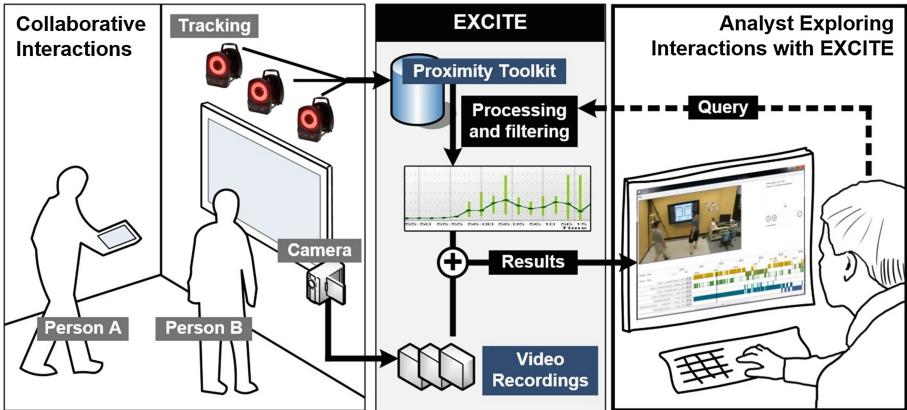
**Abstract.** A central issue in designing collaborative multi-surface environments is evaluating the interaction techniques, tools, and applications that we design. We often analyse data from studies using inductive video analysis, but the volume of data makes this a time-consuming process. We designed EXCITE, which gives analysts the ability to analyse studies by quickly querying aspects of people's interactions with applications and devices around them using a declarative programmatic syntax. These queries provide simple, immediate visual access to matching incidents in the interaction stream, video data, and motion-capture data. The query language filters the volume of data that needs to be reviewed based on criteria such as application events, and proxemics events, such as distance or orientation between people and devices. This general approach allows analysts to provisionally develop theories about the use of multi-surface environments, and to evaluate them rapidly through video-based evidence.

**Keywords:** Interaction analysis · Collaborative interaction · Tracked environments

## 1 Introduction

One important concern in designing and building multi-surface environments is ensuring that the tools and interaction techniques meet the collaboration needs of people in the environment. Researchers conduct studies of collaborative activity to understand the effect of interaction techniques and applications [9], and one of the main challenges is analysing these studies. Typically, there are many interactions in a study of collaborative behaviour in a multi-surface environment: collaborators are working with one another, or alone; they may be making use of tablets, tables, or large displays; they may be studying something on a tablet before looking up or exploring data on a different display. Yet, to determine which factors are affecting what behaviours is time-consuming. Proper analysis of this data involves time-consuming transcription and annotation of video data recorded of these studies to understand the interactions between the moving entities (e.g. [4]).

To support this analytic task, we designed EXCITE, a tool that can ease the burden of video and tracking data analysis for multi-surface environments (Fig. 1). EXCITE leverages proxemic information—such as people's and devices' distance and orientation,



**Fig. 1.** EXCITE overview: Capturing video, tracking data, and application events (center) in collaborative environments (left), and providing querying interface and visualizations to allow analysis of group interactions (right).

captured with the Proximity Toolkit [5]—simultaneously with video data and event data from the tools, and allows video analysts to generate queries on the data as a whole. These queries are annotated on a timeline, allowing the analyst to scrub through synchronised video capture feeds to validate and further annotate study sessions. Using our tool allows the analyst to compare the incidence of various events with one another, and see these in captured video of behaviour from study sessions.

Figure 1 illustrates the core features of EXCITE. It captures and synchronizes spatial tracking data (e.g., people’s and devices position, orientation, and movements) with video streams and supports interactive queries that filter these data streams. As illustrated in Fig. 2, these queries result in an annotated timeline of study sessions, where EXCITE provides a visual interface with event-scrubbers allowing an analyst to skip between occurrences of events, and even compound constructions of events. The queries themselves are constructed through a mix of declarative/imperative semantics. Because these can be constructed and visualised quickly, an analyst can rapidly explore and iteratively test different hypotheses. This kind of approach supports the style of inductive analysis that is used in video analysis [4, 6, 8, 9].

## 2 Related Work

We review tool support for facilitating qualitative analysis of interactions, in particular (1) studying interactions with video analysis, (2) evaluating prototype hardware, and (3) investigating interactions in multi-surface environments.

To facilitate the laborious task of analysing video data, tools have been developed for easier video review, motion and frame-by-frame analysis, or adding of annotations [2]. Recent systems began augmenting video data with visualizations of captured sensor information to facilitate the analysis of interactions. For example, VACA

provides a simultaneous review of video and sensor data [1]. It uses a synchronised timeline and side-by-side playback of video and additional captured sensor data. The person analysing the video can then use the provided sensor data as additional cues for finding relevant parts of the recorded interactions in the video stream.

Further specialised tools have been designed for investigating interactions with novel hardware prototypes. D.tools introduced a statechart-based analysis tool linking videos of test sessions, interaction states and other events in combined visualizations and a common event timeline [3]. ChronoViz uses multiple streams of time-based data for studying paper-based digital pen annotations [10]. Our design of EXCITE is inspired by this work, translating a similar visual inspection tool of video+sensor data to ubiquitous computing (ubicomp) interactions.

More recently, tools facilitating the analysis of multi-person and/or multi-device interactions have emerged. With pure capture of people's actions with multi-device software, VICPAM provides timeline visualizations of groupware interactions [7]. VisTACO [8] records interaction sequences of multiple remote-located people with digital tabletops, providing touch trace visualizations, and allowing insights into the spatial and temporal distribution of tabletop interaction [9]. Panoramic is an evaluation tool for ubicomp applications, providing historical sensor data visualizations [11]. Finally, the Proximity Toolkit allows recording and playback of tracked interaction sequences of multi-entity relationships (multiple people and devices) [5].

### 3 Design Rationale

In general, the prior work provides great insight into the challenges of analysing collaborative interaction, and fit closely with our own experiences with this task. The challenges of studying collaborative behaviour with technology are well documented—briefly, that in contrast to studying one user interacting with a computer (where there is only one relationship to observe), the multi-device, multi-person nature of multi-surface collaboration means there are many more relationships to be studying, observing, and to be made sense of. Researchers have generally found this challenging to do using traditional field notes, and so relying on video to enable replaying and reviewing of the interactions is common. Nevertheless, the task remains time-consuming: the common method of using an inductive, qualitative analytic approach means that a researcher needs to be able to generate hypotheses, and explore this to see whether it seems to be happening across the captured video [4].

As we have seen, sensor data can help, but only to a limited extent—sensor data provides cues into the data, but at very low levels, this sensor data can be misleading, cuing us to the wrong events in the video stream entirely (e.g. [8]). For instance, just because two study participants in a collaborative task are nearby one another, does not mean they are working with one another. But, if we also see application events at this time that indicate cross-device information transfer is happening, then we can be more certain they are working together. The main analytic challenge—making sense of what is happening (or what happened)—remains.

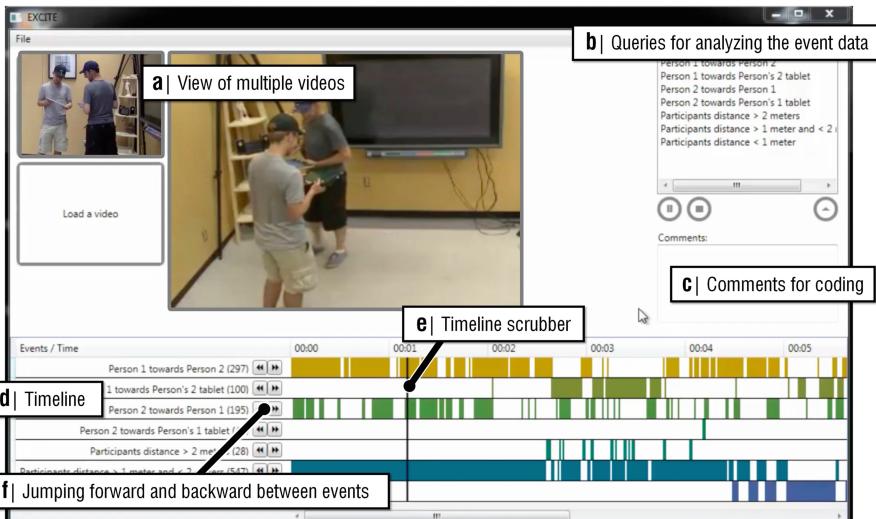
The rationale for designing EXCITE begins from this argument. To support analysis of multi-person interactions in multi-device environments, we saw that we

needed not only to integrate multiple streams of data (i.e. video data, sensor data, and also application data), but also to provide means for analysts to provisionally query the data. These queries, generated by the analyst, and executed by the system, would then allow the analyst to focus efforts on the higher-level analytic task (e.g. is collaboration happening), rather than on the low-level task of simply finding video evidence to support or refute hypotheses. Thus, beyond functioning just as a filtering mechanism, the queries allow the analyst to engage in true exploration of the data on analytic, rather than on sensor-value terms.

## 4 EXCITE: Overview and Design

We contribute EXCITE, a tool that allows rapid review of captured tracking data of interaction sessions in multi-surface environments. EXCITE facilitates the analysis of group interaction tasks in ubicomp environments, by providing an expressive querying interface and appropriate event visualizations. It unifies the access to multiple data sources (including up to four video streams and spatial tracking data).

The user interface of EXCITE consists of three major elements. First, the views of one or multiple recorded video streams of the interaction sequence (Fig. 2a). A user can add up to four video files to the viewer that can be rearranged and resized. Second, Fig. 2b shows the list of all currently entered queries for analysing the interaction. Figure 2c also shows a text box for adding comments during video coding. Finally, Fig. 2d illustrates the navigation timeline, which includes a list of all queries, including a visualization along the x-axis of the timeline indicating when the conditions of the query are met. The timeline includes a temporal navigation scrubber (Fig. 2e), and allows scrubbing the timeline to navigate forward or backwards. Alternatively, the



**Fig. 2.** User interface of EXCITE (see text for descriptions of interface elements).



**Fig. 3.** Analysis generated through queries to the EXCITE tool.

analyst can use the navigation buttons (Fig. 2f) to jump forward and backward between events.

Internally EXCITE connects to the Proximity Toolkit [5] for capturing the tracking information of people's and devices position and orientation. The Proximity Toolkit uses depth-sensing cameras to track people's location, and a high end motion capturing system with infrared-reflective markers to track devices. EXCITE records the proxemic information provided by the toolkit—that is, distance, orientation, identities, movement, and location—and stores this data in an internal data structure to perform the queries.

## 5 Query Language

The query language is the core toolset provided by EXCITE for analysing the recorded video and sensor data of the performed interaction. The queries allow filtering and analysing the captured data for quickly finding particular events that happened during interaction.

### 5.1 Structure and Composition of Queries

Each query is composed out of (1) one or two presence/application-event identifiers, (2) a function to compare or a property to check, and (3) a condition. An analyst can add as many individual or compound queries for the event stream as they need; any new query is added as a new horizontal parallel event stream in the timeline view (Figs. 2d and 3).

The *presence identifiers* directly correspond to the identification names of entities tracked in the Proximity Toolkit. These can be identifiers for people such as 'Person1' or their name such as 'Taylor', or for devices such as 'Smartboard' (for the large interactive surface) and 'Tablet' (for an interactive tablet computer). EXCITE also handles *application event identifiers*, where the system can read log files generated by applications (e.g. Tablet1.TouchDown).

The next components are the *functions*. They can compare values between two entities or check a property of a single entity. The available functions (with Boolean or Integer return values) for comparing two presence identifiers are:

- **Distance (Integer):** distance between entities (in mm).
- **Velocity\_difference (Integer):** the velocity difference between two tracked entities.
- **Orientation\_difference (Integer):** the difference in orientation angles of the default pointer in the proximity toolkit (in degrees).
- **Towards (Boolean):** true if an entity is pointing towards another entity (orientation angles divided in two sections at +90 to -90).
- **Pointing\_at (Boolean):** true if an entity's previously defined pointing vector (e.g., the normal vector of a screen) is directly pointing at another entity.
- **Touching/Colliding (Boolean):** true if two entities are either touching or their bounding volumes are colliding (below a set fixed threshold).
- **Parallel/Perpendicular (Boolean):** true if default pointing vectors are parallel or perpendicular.

Each function can be combined with the names of *two tracked entities* and a *condition* to compare to. Valid operators for the comparison are:  $\leq$ ,  $<$ ,  $>$ ,  $\geq$ ,  $=$ , and  $\neq$ . For example, the following query checks if the distance between two people is smaller than 1 m (1000 mm):

```
person1.distance(person2) < 1000
```

As another example, a query can check if a person is facing towards the large display (i.e., the smartboard) or facing away (using the *towards* function):

```
person1.towards(smartboard)
```

Individual properties of an entity include the entity's 3D coordinates (X, Y, Z values) as well as orientation and velocity values. Again, the properties can be combined with a condition to filter only the events of interest.

## 5.2 Compound Queries and Parallel Event Streams

Combining multiple conditions into compound queries gives analysts a powerful tool for refining the hypothesis investigate with EXCITE. Multiple statements can be combined with logical operators (`&&` for *and*, `||` for *or*, `!` for *not*), and compound queries can be composed of any number queries and logical operators. The following example's query only returns results if both concatenated individual queries are valid: first, if the distance between two people is smaller than 1 m, and second, if 'person1' is facing towards the large display.

```
person1.distance(person2) < 1000 && person1.towards(smartboard)
```

Because the queries can be constructed and visualised quickly, the analyst can rapidly generate, explore and refine different hypotheses. This kind of approach supports the inductive analytic methods used by video analysts.

## 6 Analysis Walkthrough Case Study

We now show a case study example to demonstrate how an analyst can apply EXCITE in practice when analysing group interactions with interactive surfaces.

Larry has designed a simple multi-surface system that allows people to share information to others via two interaction techniques (inspired by [6]): (1) *portals*—if tablets are held close to one another, information can be swiped from one display to the next, and (2) *hold-to-mirror*—if the tablet is turned to face toward the large display, then information on the tablet is transferred onto the large display. Larry is interested in how people share information in his study.

After running his study, Larry loads the logged data streams from his application, which populates on the timeline as two separate tracks: one for “portal” transfers, and one for “hold-to-mirror” transfers (Fig. 3a). As Larry goes through the synchronised video data, he notices that the “hold-to-mirror” events match up with times when participants are not close together. He generates the following query:

```
person1.distance(person2) > 4000
→ When are the people far away from each other (more than 4m)?
```

This does not seem to return many query results—participants stayed within 4 m of each other most of the time. He *refines* the query with a smaller window:

```
person1.distance(person2) > 3000
→ When are the people further than 3m from one another?
```

Based on the new track (Fig. 3b), he is able to see that indeed, every one of the “hold-to-mirror” transfers happens when participants are not standing too close to one another (Fig. 3b). Out of curiosity, Larry refines his query again to see whether close distances correspond with “portal” transfers:

```
tablet1.orientation_difference(tablet2) < 15 && person1.distance(person2) < 1000
→ When are tablets oriented in the same direction, AND people are close together?
```

The results of this compound query do correspond with most of the “portal” transfers, yet the window in which this is happening seems quite big (Fig. 3c). As Larry goes each incident by inspecting the video, Larry realises that participants are actually sharing information by showing each other their tablets rather than strictly using the “portal” tool! He constructs a final query to capture this (Fig. 3d):

```
person1.distance(tablet2)< 500 || person2.distance(tablet1)< 500
→ When are people close to the other person's tablet?
```

While the result of this query is not perfect, Larry can use the results to cue him to parts of the video where one participant might be sharing information with another by simply showing the tablet to another person. Without EXCITE, Larry would be left to

review the video data, perhaps using a manual annotation tool, and not arrive at his final theory until much later in the process.

## 7 Conclusion

**Opportunities and Limitations.** While EXCITE is a flexible tool facilitating analysis of collaborative interactions, it is a proof-of-concept system. The current lexical power is limited to the query language outlined in Sect. 5, though there is ample opportunity to extend this: for instance, by allowing analysts to script altogether new semantics (e.g. walking, turn-around). These functions, which might take into account proxemic variables over time, could then be used for the filtering of data. Similarly, the UI can be improved to allow for dynamic filtering—for instance, rather than entering a specific value for a velocity or orientation, one would be able to manipulate a slider. The current implementation also depends on the Proximity Toolkit, which mainly functions with the high-end VICON and OptiTrack motion capturing systems. While in principle the Proximity Toolkit can be extended to other position/orientation capture systems, this dependency limits the current application of EXCITE.

**Summary and Conclusion.** We contributed the design of a novel tool for analysing interaction sequences in multi-person and multi-device environments. The EXCITE tool and its query language facilitate the rapid inspecting of time-synced video and captured motion-tracking data—and support finding answers to enquiries about participants' use of ubicomp gestural interactions with tablets, walls, and tabletops. In our own work, we have begun actively exploring the use of EXCITE to support our analysis, and are looking for ways of both improving the power of the query language, as well as simplifying the syntax. We are interested in also exploring how spatial semantics (e.g. [8]) can be integrated into textual queries of the data. We demonstrated the potential of these kinds of analysis with our walkthrough, and believe that the design of EXCITE will be valuable for future studies investigating people's interactions in ubicomp ecologies (EXCITE is available for download at: <http://grouplab.cpsc.ucalgary.ca/cookbook/index.php/Toolkits/EXCITE>).

## References

1. Burr, B.: VACA: a tool for qualitative video analysis. In: CHI 2006 Extended Abstracts. ACM (2006)
2. Dasiopoulou, S., et al.: A survey of semantic image and video annotation tools. In: Knowledge-Driven Multim. Information Extraction and Ontology Evolution. Springer (2011)
3. Hartmann, B., Klemmer, S., et al.: Reflective physical prototyping through integrated design, test, and analysis. In: Proceedings of UIST. ACM (2006)
4. Jordan, B., Henderson, A.: Interaction analysis: foundations and practice. J. Learn. Sci. 4(1), 39–103 (1995)

5. Marquardt, N., Diaz-Marino, R., et al.: The proximity toolkit: prototyping proxemic interactions in ubiquitous computing ecologies. In: Proceedings of UIST 2011. ACM (2011)
6. Marquardt, N., Hinckley, K., Greenberg, S.: Cross-device interaction via micro-mobility and F-formations. In: Proceedings of UIST 2012, pp. 13–22. ACM (2012)
7. Moghaddam, R.Z., Bailey, B.: VICPAM: a visualization tool for examining interaction data in multiple display environments. In: Smith, M.J., Salvendy, G. (eds.) HCII 2011, Part I. LNCS, vol. 6771, pp. 278–287. Springer, Heidelberg (2011)
8. Tang, A., Pahud, M., Carpendale, S., Buxton, B.: VistACO: visualizing tabletop collaboration. In: Proceedings of ITS 2010, pp. 29–38. ACM (2010)
9. Tang, A., et al.: Three's Company: understanding communication channels in three-way distributed collaboration. In: Proceedings of CSCW, pp. 271–280. ACM (2010)
10. Weibel, N., Fouse, A., Hutchins, E., Hollan, J.D.: Supporting an integrated paper-digital workflow for observational research. In: IUI. ACM (2011)
11. Welbourne, E., Balazinska, M., Borriello, G., Fogarty, J.: Specification and verification of complex location events with panoramic. In: Flóréen, P., Krüger, A., Spasojevic, M. (eds.) Pervasive 2010. LNCS, vol. 6030, pp. 57–75. Springer, Heidelberg (2010)

# The Usefulness of Method-Resources for Evaluating a Collaborative Training Simulator

Ebba Thora Hvannberg<sup>(✉)</sup>, Gyda Halldorsdottir, and Jan Rudinsky

University of Iceland, Dunhaga 5, 107 Reykjavík, Iceland  
`{ebba,gyda,janr}@hi.is`

**Abstract.** Voice communication is vital for collaboration between first responders and commanders during crisis management. To decrease cost, training can take place in a virtual environment instead of in a real one. It is non-trivial to build and evaluate a virtual environment for training complex command. To understand the method-resources required for evaluating a training simulator for crisis response, this paper presents a case study of applying several resources. Method-resources were analysed for usability problems and Mechanics of Collaboration (MOC). The results show that the Group Observational Technique and the MOC analysis are appropriate for analysing factors of collaboration and communication. The think-aloud technique, observers, experts in the domain and advanced task scenario were important resources. In only a few cases sound and video were necessary to analyse issues.

**Keywords:** Virtual reality · Collaboration · Evaluation · Crisis management · Verbal communication · Method-resources · Mechanics of collaboration

## 1 Introduction

Technological developments of collaborative systems have increased demands on resources for evaluating virtual environments that allow multiple modes such as visualization, verbal communication and sound. Evaluating usability of interactive systems designed for Computer Supported Cooperative Work (CSCW) has been found challenging. Because of collaborating users, the evaluation requires more observers than in a single user scenario [1]. Increasingly, sound plays a large role in virtual environments in the form of natural or synthetic sound, music or voice communication [2]. Evaluation of the effect of sound on voice communication has been carried out but mostly in controlled laboratory settings [3].

To gain valid results of a usability evaluation, method selection continues to be a critical activity. The increased complexity of systems has called for ways to choose methods that are most appropriate for usability evaluations. One such framework has been proposed by Antunes et al. [4] who devised a selection strategy for choosing an appropriate method of evaluation. Unfortunately, empirical research on usability evaluation of collaborative virtual environment has not been extensive [4]. While evaluations of collaborative environments have been mostly studied theoretically [5], studies have been reported on the effectiveness of 3D collaborative virtual environments [6]. Recognizing that multiple

dimensions of interactive systems call for more than an adoption of one or two evaluation methods, it has been suggested that a collection of resources, that are a part of a whole method, return an efficient evaluation [7]. Examples of resources of methods are participant recruitment, task scenario, reporting format, problem identification, problem classification and a thinking-aloud protocol [7]. This paper uses the term method-resources to describe such resources [8].

Thus motivated, the main objective of this research is to gain empirical knowledge on the suitability of resources for usability evaluation of a collaborative system, set in a virtual environment, where players communicate verbally in a noisy environment. To achieve this objective, we studied a usability evaluation of a prototype virtual environment for training crisis management personnel dealing with mass-casualty accidents.

Crisis management training is organized and developed according to an accurate predefined system for immediate responses to mass-casualty incidents. First responders are personnel from volunteers to professionals with a few to many years of experience. Professional responders come from different organizations, e.g., rescue, police, medical and firefighting. Training these people to obtain efficient skills is crucial.

The application (see Fig. 1) that was evaluated is a virtual training simulator for crisis response to a mass-casualty incident at an airport. Trainees are commanders, the On Scene Commander (OSC) managing the responses overall on the scene and the Rescue Coordinator (RC) managing the rescue work on the scene that exercise progress report, rescue resources, information requests and task delegation. They can navigate with an on scene view and a zooming in effect through the virtual environment that has a changing high-fidelity scene. They are wearing earphones with a microphone. A design based on empirical data consisted of three voice communication metaphors reflecting the communication spaces used in a real crisis event and its training, i.e. one way radio, mobile phone and face-to-face (F2F) communication for the two persons to speak to one another [9]. To broadcast a message a GUI button is pressed with a mouse (i.e. Push-to-talk) and in response, a transfer is reversed from receiving to sending with the player microphone no longer muted. The prototype has a high fidelity soundscape with sound from fire trucks, other resources, players talking, fire and wind. Any type of sound is relayed over communication channels, i.e. players can hear noises at the far end through the channel.



**Fig. 1.** A snapshot of the virtual environment showing the scene of the accident.

## 2 Selection of Method-Resources

### 2.1 Evaluation Methods

To select an evaluation method we applied the CSCW evaluation framework by Antunes et al. [4], where a variety of evaluation methods are presented for each stage of the software development. We selected Groupware Observational User Testing (GOT) as an end-user oriented evaluation method focusing on realism and usability as the main objective [4, 10]. Gutwin and Greenberg [10] proposed GOT, as a cost effective usability method, which was based on a set of fundamental Mechanics Of Collaboration (MOC). The GOT technique is an observational user testing method focusing on usability in a planned situation, collaboration where users perform predefined tasks. The framework of MOC includes seven categories of important collaboration activities: Communication as Explicit communication and Consequential communication (information unintentionally given off by others); Coordination of action; Planning; Monitoring and gathering information in the workspace; Assistance to one another; and Protection of resources in the workspace [10, 11]. The MOC model has been evolving [12], but the original set was more appropriate for our study. Besides MOC, the method-resources needed for GOT are think-aloud, observers, users and tasks. Furthermore, to record observations, screen-captures were used and audio recorded of the communication in the virtual environment and of the think-aloud. The tasks and the users are described in the next section.

### 2.2 Collaborative Scenario and Users

Based on extensive observations of crisis management training on-site, interviews and workshops, a collaborative scenario comprising several tasks was written and validated by an experienced crisis management instructor. The scenario aimed to secure a situation at an accident scene and allowing commanders to ask for resources, such as fire fighters, using verbal communication (see Table 1). Each commander was located in a separate room in front of a screen wearing a head-set with a microphone.

Six employees of a rescue and fire organisation at an airport with experience in crisis response were recruited as participants for the study and divided into three pairs of OSC and RC. Participants had all received a one day introduction to training in a virtual environment, but not to the prototype used in this study. The same session was repeated three times, once for each pair of collaborators which were followed by an observer.

### 2.3 Analysis of Usability Problems and Protocol Analysis According to MOC

The data was analysed in two ways, analysing usability issues and collaborations using MOC. A bottom up qualitative analysis was performed identifying usability issues that were consolidated into unique usability problems. Before analysing the data, the second author transcribed the audio data into text while listening to it and observing the video capture. Comments that observers (the first and third authors) had written down during the sessions were integrated to the transcript protocol. The third author analysed the protocol for problems that participants faced that were then verified by the first author. In addition to problems, an observer looked for successful interactions, activities and comments raised

**Table 1.** Tasks in a collaborative scenario for OSC and RC.

On Scene Commander tasks	Rescue Coordinator tasks
1. When receiving a mobile call with incident details, please note down the information.	
2. Set up the emergency channel on your radio device by configuring to 116Hz frequency.	1. Set up the emergency channel on your radio by configuring it to 116 Hz frequency.
3. You must navigate to the gate. If a person later appears at the gate you will ask him for a name and register the name by writing it down on paper.	
4. You can use radio or mobile phone to contact the RC. Inform RC of the incident details that you received previously and ask him or her to report at the gate.	2. Wait for a message from OSC on radio or mobile phone. Once asked by the OSC you will go to the gate.
	3. When you reach the gate you will notify the OSC of your presence and state your name.
	4. Go to the crashed plane and count casualties.
	5. Give the number of casualties to the OSC using radio.
5. You will coordinate the rescue operation with RC. If you receive important information you should write it down. If you receive a call from the RC asking for additional resources, locate the required resources around and send them to the scene.	6. Observe the development of fire and listen to messages. If the fire spreads covering most of the plane, contact OSC for additional resources.
6. Once the fire is out, you can go to the scene. If you are contacted by RC asking about casualties' placement, you suggest a location that is near the plane, but safe from fire, smoke and flying debris and not in the line between the scene and gate.	7. If you receive a report by the fire fighter's team leader that the fire has been extinguished, you must inform OSC that it is now safe to enter the scene. OSC can be contacted over the radio or F2F if he or she is around.
	8. Discuss casualty placement with OSC over radio or F2F and make an agreement on possible locations.

by the participants. After analysing the protocol, the observations were categorised into groups emerging from the data. Reviewing the transcribed conversations and the videos of the screen capture, the second author and an engineer analysed the data using MOC. After analysing them independently, they discussed differences and came to a consensus. Additional method-resources are observers with expertise in human-computer interaction and moderate expertise in the domain of crisis management.

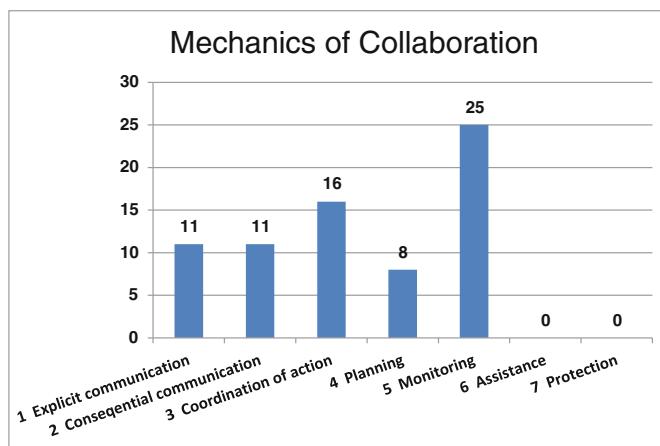
### 3 Results

The experience gained from applying the method-resources will help analyse their usefulness. The sessions lasted 17–20 min each, which gave 109 min of transcribed audio and video recordings. The number of conversations per pair was 10–12, or a total of 32 for all sessions. Observations, capturing problems or successful interactions, e.g. successful training, were 112.

The data analysis uncovered 13 unique usability problems of 84 problem instances divided into eight categories, communication, information/communication, collaboration, navigation, discrepancies between virtual and real world, sound, wrong or inappropriate

tasks and following scripts. The first three problem categories of communication and collaboration are discussed in this paper.

Altogether 71 collaboration instances from the conversations of the three pairs of participants were analysed with respect to the MOC categories (see Fig. 2). It is noteworthy that the players seemed to monitor their environment and the situation extensively, learning where people are and what they are doing. That no assistance took place could be attributed to low complexity of the scenario and that the players saw no need to protect their workspaces indicates that there were no threats imposed on them in the environment. To analyse the collaboration protocol according to MOC an exact transcript of the collaboration was required.



**Fig. 2.** Collaborations analysed using MOC.

The method-resources proved to be helpful for uncovering issues and usability problems of categories addressed here. Table 2 shows examples of how the method-resources were used to uncover particular issues. The issues in the table are labelled with issue IDs (e.g. A01) which are used in the following text and the numbers of the MOC categories are identified according to Fig. 2.

To show how the different method-resources have helped, a few examples of the collaboration protocol are given. In several cases, we noticed that the trainees were able to practice skills successfully. One of them is a coordination mechanic (A01):

*OSC: [Thinking aloud] I will meet him. The wind actually is like that. I will call. Oops. I will go and meet him like, somewhere here. Not too close to the fire though. I would meet him around here.*

The Think-aloud protocol helped the observer to gain insight into what the user is thinking and discovered that he was practicing skills successfully. Such a scenario could not have been practiced without expert users in the domain.

In two of the three experiments, it took participants some time to realise that they should collaborate and take turns, but after a short while they got used to it (A02). Four usability problem instances were observed affecting two participants:

**Table 2.** Examples of use of method-resources.

Issue ID	Description	Method-Resources										MOC categories
		Expert in domain	Task scenario	Thinking aloud protocol	Voice	Sound	Video	Observation	Notes by observers	Transcripts of collaboration	Problem identification	
A01	Thinking aloud; Checking conditions	X	X	X	X				X	X	X	3
A02	Thinking aloud; Checking ones role in the experiment;	X	X	X	X		X	X	X	X	X	3
A03	Acknowledging information wrongly.	X	X		X	X	X	X	X	X	X	1
A04	OSC is not aware of that RC can hear him. Radio is up on the screen, but it is not lighted as put on for talking			X	X		X		X	X	X	2
A05	Needing some guidance to navigate			X	X		X		X	X	X	3
A06	Does not realize that he is located at wrong place			X	X		X	X	X	X	X	5
A07	Mismatch in prototype vs. real	X	X	X	X		X	X	X	X	X	5
A08	RC can't hear everything OSC is saying, although OSC is not aware of			X	X		X	X	X	X	X	5
A09	Coordinating own action. Needing guidance to call.	X	X	X	X		X		X	X	X	5

*OSC: [Thinking-aloud] So, Do I then... – do I wait for him to call a backup team or help. Or do I? Is that my decision, or does he...*

The current method-resources make it difficult to conclude the nature of this problem and the cause of the confusion. It could be that the commander is unsure of the collaborative scenario, that the virtual environment does not provide adequate affordance, or even that he has not been trained adequately in his role. An additional method-resource would be needed to inquire about certain critical points.

Some problems were about providing information verbally. Users relayed wrong information to their partners, or missed to respond when being addressed, either altogether or not responding accurately (A03):

*OSC: “RC we have a plane crash on runway 19 the intersection. Plane is on fire. We have 35 people on board. It is a mini-jumbo jet. You go on scene with your team.”*

*RC: “RC got that. A plane crashed down on fire, an intersection 11. No 19 and 01. And 38 people on board” “What type of aircraft is that?”*

The participant acknowledged inaccurately and wrongly, or he did not hear correctly the number of people on board, i.e., 38 instead of 35. The video was checked to confirm conditions that showed good sound and no noise disturbing. Five participants had such a problem in a total of 15 instances. It is not analysed as a usability problem and it may even reflect an accurate picture of a normal training scenario. Another thing we noticed is that the expert users in the domain were able to play their role and act out from the given scenario (“What type of aircraft is that?”), thus indicating that the virtual environment is a useful training tool. The MOC analysis proved useful in separating the explicit communication issues from others. The advanced collaborative scenario and expert users in the domain are vital method-resources to create a dialogue and reveal such a scenario.

It was observed that OSC talked while navigating to the gate in the virtual environment and kept the radio channel open, allowing RC to hear what he said (A04, A05,

A06, and A07). RC told the observer that he could hear everything OSC said (A08). Later, OSC resolved how to use the radio and called his partner successfully by identifying himself and the receiver (A09). We noticed that participants used the phone much less but encountered similar problems. Two of the three pairs tried to talk F2F, one of them took a few minutes to make it work smoothly but a second pair used it without problems. Expertise of the users in using communication devices in crisis management, e.g. radios, was crucial to understand how they used the communication metaphors in the virtual environment. Observers are expensive resources for evaluating how much certain features are used and could be replaced with automatic monitors. Finally, we see in this example, when noting that the OSC talked on his way to the gate, that it is essential to have a screen-capture of the experiment.

The GOT method and its resources worked well for focusing on usability evaluation in a planned situation, especially when focusing on collaboration and users performing particular predefined tasks. The factors of collaboration and communication were the focus of this study and using the MOC analysis of collaborations fits well for them. The examples show that observation, transcripts and the think-aloud technique are fundamental in researching verbal communication in a collaborative environment. Experts in the domain and the task scenarios are used to uncover fewer issues, but are nonetheless essential. The captured voice is used to produce exact transcript of the collaboration, but as Table 2 shows the sound and the video is used less to understand the issues.

## 4 Conclusion

The contribution of the paper is twofold. First, we have shown a method for analysing the usefulness of method-resources. Such a method can be useful for other researchers analysing method-resources. Second, its application shows that the GOT method and the MOC analysis are appropriate for analysing factors of collaboration and communication. An essential part of that is to include the think-aloud technique and observers. In only a few cases sound and video were necessary to analyse issues. Other resources that were especially important were expert users in the domain and advanced task scenario.

A few ideas emerged for decreasing the cost of method-resources and raising their effectiveness. Developing software tools for monitoring the scenario, e.g. the frequency of use of features could help decrease the expenses of observation in the collaborative scenario and data analysis. In an environment where the dialogue is rich, the domain is complex and user domain expertise is high, it may be more difficult than ever to understand the causes of users' actions. A method-resource to analyse critical points of understanding causes may be needed.

## References

1. Grudin, J.: Why CSCW applications fail: problems in the design and evaluation of organizational interfaces. In: Proceedings of the 1988 ACM Conference on Computer-Supported Cooperative Work, pp. 85–93. ACM, Portland (1988)

2. Divjak, M., Kore, D.: Visual and audio communication between visitors of virtual worlds. In: Advances in Neural Networks and Applications, pp. 41–46 (2001)
3. MacDonald, J.A., Balakrishnan, J., Orosz, M.D., Karplus, W.J.: Intelligibility of speech in a virtual 3-D environment. *Hum. Factors: J. Hum. Factors Ergon. Soc.* **44**, 272–286 (2002)
4. Antunes, P., Herskovic, V., Ochoa, S.F., Pino, J.A.: Structuring dimensions for collaborative systems evaluation. *ACM Comput. Surv.* **44**, 1–28 (2012)
5. de Freitas, S., Rebollo-Mendez, G., Liarokapis, F., Magoulas, G., Poulovassilis, A.: Developing an evaluation methodology for immersive learning experiences in a virtual world. In: Games and Virtual Worlds for Serious Applications, VS-GAMES 2009. Conference in, pp. 43–50 (Year)
6. Montoya, M.M., Massey, A.P., Lockwood, N.S.: 3D collaborative virtual environments: exploring the link between collaborative behaviors and team performance. *Decis. Sci.* **42**, 451–476 (2011)
7. Woolrych, A., Hornbæk, K., Frøkjær, E., Cockton, G.: Ingredients and meals rather than recipes: a proposal for research that does not treat usability evaluation methods as indivisible wholes. *Int. J. Hum.-Comput. Interact.* **27**, 940–970 (2011)
8. Law, E.L.-C., Hvannberg, E.T., Vermeeren, A.P., Cockton, G., Jokela, T.: Made for sharing: HCI stories of transfer, triumph and tragedy. In: CHI 2013 Extended Abstracts on Human Factors in Computing Systems, pp. 3235–3238. ACM (Year)
9. Rudinsky, J., Hvannberg, E.T., Helgason, A.A., Petursson, P.B.: Designing soundscapes of virtual environments for crisis management training. In: Proceedings of the Designing Interactive Systems Conference, pp. 689–692. ACM, Newcastle Upon Tyne (2012)
10. Gutwin, C., Greenberg, S.: The mechanics of collaboration: developing low cost usability evaluation methods for shared workspaces. In: Proceedings of the IEEE 9th International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises, (WET ICE 2000), pp. 98–103 (Year)
11. Pinelle, D., Gutwin, C.: Evaluating teamwork support in tabletop groupware applications using collaboration usability analysis. *Pers. Ubiquit. Comput.* **12**, 237–254 (2008)
12. Pinelle, D., Gutwin, C., Greenberg, S.: Task analysis for groupware usability evaluation: Modeling shared-workspace tasks with the mechanics of collaboration. *ACM Trans. Comput.-Hum. Interact.* **10**, 281–311 (2003)

# Flat Design vs Traditional Design: Comparative Experimental Study

Ivan Burmistrov<sup>1,2()</sup>, Tatiana Zlokazova<sup>1</sup>, Anna Izmalkova<sup>1</sup>, and Anna Leonova<sup>1</sup>

<sup>1</sup> Laboratory of Work Psychology, Lomonosov Moscow State University, Moscow, Russia

{t.zlokazova,ableonova}@gmail.com, mayoran@mail.ru

<sup>2</sup> InterUX Usability Engineering Studio, Tallinn, Estonia

ivan@interux.com

**Abstract.** In the past few years flat user interface design has become the predominating visual style of operating systems, websites and mobile apps. Although flat design has been widely criticized by HCI and usability experts, empirical research on flat design is still scarce. We present the results of an experimental comparative study of visual search effectiveness on traditional and flat designs. The following types of visual search tasks were examined: (1) search for a target word in text; (2) search for a target icon in a matrix of icons; (3) search for clickable objects on webpages. Time and accuracy parameters of the visual search, as well as oculomotor activity, were measured. The results show that a search in flat text mode (compared with the traditional mode) is associated with higher cognitive load. A search for flat icons takes twice as long as for realistic icons and is also characterized by higher cognitive load. Identifying clickable objects on flat web pages requires more time and is characterised by a significantly greater number of errors. Our results suggest replacing the flat style user interfaces with interfaces based on the design principles developed over decades of research and practice of HCI and usability engineering.

**Keywords:** Flat design · Usability · Visual search · Cognitive load · Eye-tracking

## 1 Introduction

In 2012–2014 the design of user interfaces for the operating systems (OS) of desktop computers, mobile OS and mobile applications, as well as for websites, saw cardinal changes relating to the appearance of so-called flat user interface design. The first flat design appeared in the mobile OS Windows Phone 7 in 2010. It came to prominence two years later with the OS Windows 8 for personal computers. This new approach to the design of user interfaces was enthusiastically received by the graphic design community as well as by many users, as a result of which it was adopted by two other leading software vendors, Apple and Google.

The basic flat design principle means that the computer screen represents a self-contained two-dimensional digital environment in which there is no place for anything replicating three-dimensional objects of the real world [2]. The user interface elements are simplified: abstract graphic forms are used and spaces are filled with bold colours.

Text and font are especially important in flat design. In particular, this leads to a wide use of condensed, light and ultralight variations of typefaces. The density of screen information is often extraordinarily low [10].

Shortly after its introduction, flat design became subject to criticism by HCI and usability experts [3, 6, 10–12, 18]. The main criticism was that flat design ignores the three-dimensional nature of the human brain, which is extremely sensitive to visual cues linking interfaces to the real world. The removal of affordances from interactive interface objects means that users regularly perceive interactive elements as non-interactive, and non-interactive elements as interactive.

Despite these limitations flat design is becoming more and more common, and criticism of experts in HCI and usability is generally ignored by the software industry and graphic designers. Unlike these expert assessments (from 2012 onwards) the results of empirical research into flat design are still not numerous, so it is important to conduct more comparative experimental studies of flat and traditional design.

## 2 Background and Related Work

Recent empirical research has mostly considered a quantitative comparison of the performance measures by users of both traditional and flat style interfaces, and users' emotional reactions and preferences for realistic and flat icons.

A comparative usability study of Windows 8 (flat interface) and its predecessor Windows 7 (traditional interface) showed that Windows 7 was superior to Windows 8 in each of three aspects of usability: effectiveness, efficiency and satisfaction [15].

In research carried out by Idler [8] 100 web professionals completed tasks relating to the clickability of objects on four flat websites, and also assessed the advantages and disadvantages of flat design. The results showed that the number of 'false alarm' errors when working on flat sites varied from 16 % to 38 % (average 29 %). The authors concluded that despite the apparent clarity and simplicity of flat design, achieving an acceptable level of website usability is not easy.

Comparative research of the aesthetic perception by users of pairs of realistic and flat icons of applications for desktop computers and mobile devices indicated that the users preferred realistic icons to flat icons by a proportion of 75:25 [7]. In another study, flat icons scored higher on semantic scales such as "timeliness" and "simplicity", but they fared worse than realistic icons in "identity", "interest" and "familiarity" aspects [9]. A semiotic inspection of icons of standard applications for iOS 6 (realistic icons) and iOS 7 (flat icons) showed that the unsuccessful transformation from realistic to flat icons is often related to the loss of semantically important attributes during the "simplification" process inherent to flat design [16].

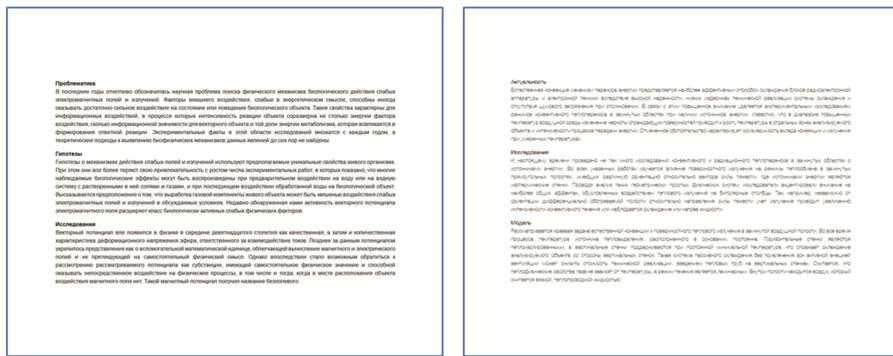
## 3 Method

As can be seen from the above-mentioned studies, an empirical analysis of flat design is still at an early stage. In our experiment we tried both to build on previous research, but also to include in our analysis several new aspects. In order to conduct an accurate

comparative study of traditional and flat interfaces we chose the following design elements: fonts, icons and webpages. In addition to the classical performance measures like time on task and number of errors, we also included an analysis of oculomotor indicators of cognitive load.

The experiment consisted of two series: traditional and flat. In each series of the experiment participants carried out three types of task:

- (1) A search for a target word on a page comprising three paragraphs, all typed using the same font (Fig. 1).



**Fig. 1.** Examples of traditional and flat text stimuli

For the traditional series we used three similar typefaces: Helvetica Neue, Arial and Tahoma (these are fonts used in older versions of OS Windows and pre-flat era websites); for flat series we used three variations of Helvetica Neue font: Condensed Normal, Light and UltraLight (condensed fonts are popular on modern websites, while Helvetica UltraLight and Light were system fonts in iOS 7, beta and final versions respectively) (Table 1). The target word (e.g. “structure”) was placed randomly in the first, second or third paragraph. The participants were instructed to click the target word.

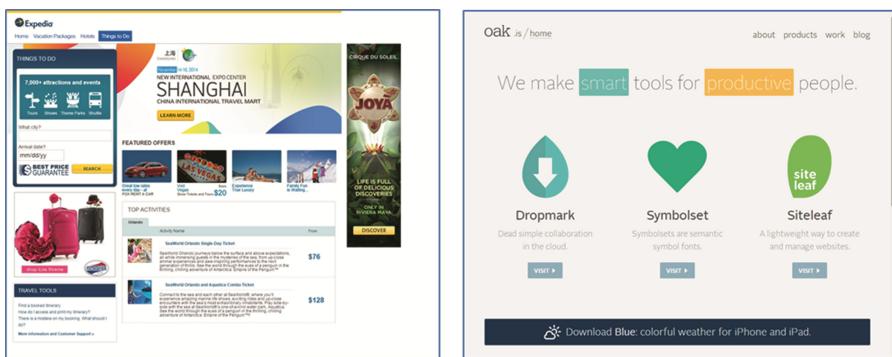
**Table 1.** Typefaces used in the experiment

'Traditional' fonts	Helvetica Neue	Arial	Tahoma
'Flat' fonts	Helvetica Neue Condensed Normal	Helvetica Neue Light	Helvetica Neue UltraLight

- (2) A search for an icon depicting a specific object (e.g. “ice-cream”) in a matrix of 9 × 9 icons presented on the screen (Fig. 2). The position of the target icon was randomly distributed between the nine quadrants of the matrix. The participants were instructed to click the target icon.
- (3) A search for clickable objects (i.e. objects which change something on the screen after a click) on screenshots of existing websites (Fig. 3). The participants were instructed to click all screen objects that look clickable (buttons, links, menus, images, banners etc.).



**Fig. 2.** Examples of realistic and flat icons stimuli



**Fig. 3.** Examples of traditional and flat web page stimuli

In each series of the experiment the participant was given 9 slides with text, 9 slides with icons and 9 webpage screenshots. The order of tasks in each series was the following: first the participant was given one text search task (each of the typesets appeared three times within a series), then one icon search task, then one screenshot search task. This pattern was then repeated until all 27 tasks had been completed. The order of the different series of the experiment (first traditional, then flat – or the other way round) was counterbalanced. Before the experiment began, the participants were given instruction and training.

The stimuli were presented on a 19 inch LCD monitor with 4:3 aspect ratio. To measure the participant's eye movements an EyeLink 1000 eye-tracker was used. All screen events and mouse movements were recorded using the TechSmith Morae 3.2 data logging application.

For the search of target words and icons the performance time was measured. For the screenshot task we measured the average time taken to click all clickable objects on a slide, and also registered the number of 'miss' and 'false alarm' errors.

For each task type the mean eye-tracking indicators were analysed, including fixation and saccade parameters. These parameters are considered in the literature as indicators

of the cognitive load and show the following dynamics when the cognitive load increases: an increase in fixation duration [4, 14], a decrease in saccadic amplitude [13, 19], and a decrease in saccadic peak velocity [1, 5].

Participants were: 19 female and 1 male university student from Moscow, aged 18–28 (mean – 21.2), experienced web, smartphone and tablet users. The experimental sessions were conducted in November 2014. By that time flat style already predominated on desktops and mobiles, and so was familiar to all the participants.

## 4 Results and Discussion

**Fonts.** Mean values of performance time and oculomotor measures for the text search task are shown in Table 2.

**Table 2.** Results for the text search

Measure	Traditional series Mean ( $\sigma$ )	Flat series Mean ( $\sigma$ )	Student $t$ -value	$p$
Performance time (sec)	46.0 (16.7)	42.7 (12.3)	1.258	0.215
Fixation duration (msec)	256 (34.6)	266 (32.9)	3.462	0.001
Saccadic amplitude (deg)	3.7 (0.81)	3.3 (0.74)	3.967	0.001
Saccadic peak velocity (deg/sec)	138 (18.7)	128 (18.6)	3.919	0.001

Statistical analysis did not reveal any significant difference in performance time. At the same time, oculomotor indicators of increased cognitive load – increase in fixation duration, decrease in saccadic amplitude and saccadic peak velocity – showed statistically significant differences in the traditional and flat series. This type of combination of measures (long fixations and short saccades) is characteristic of focal visual information processing: i.e. a conscious analysis of information, precise identification of objects and events, which are implemented when the visual search tasks increase in complexity [17]. In the text search with traditional fonts, subjects had more opportunities to switch to “semi-automatic” information processing associated with a lower cognitive load (which is indicated by shorter fixations and longer saccades). Also the lower values of saccadic peak velocity provide evidence in favour of associating the text search in the flat series with a higher cognitive load.

**Icons.** Mean values of performance time and oculomotor measures for the icon search task are shown in Table 3.

**Table 3.** Results for the icon search

Measure	Traditional series Mean ( $\sigma$ )	Flat series Mean ( $\sigma$ )	Student <i>t</i> -value	<i>p</i>
Performance time (sec)	8.0 (2.2)	15.4 (4.6)	5.611	0.000
Fixation duration (msec)	284 (70.4)	264 (58.4)	0.857	0.403
Saccadic amplitude (deg)	4.6 (1.11)	3.2 (0.93)	8.728	0.000
Saccadic peak velocity (deg/sec)	174 (25.2)	137 (22.5)	8.810	0.000

A significant difference was found in the mean values of the icon search time: almost twice as high for flat as for realistic icons. Unlike in the previous task, a comparison of oculomotor activity in the graphic objects search did not reveal any significant difference in mean fixation duration. Nevertheless, a difference in mean saccadic amplitude and saccadic peak velocity remained, just as in the text search task. Values were less in the flat series, which may indicate the higher complexity of the task and a higher cognitive load in the flat icon search.

Monitoring the performance process of this task allowed us to assume that many participants in the flat series could not find the target icon during the initial “fast” slide scanning. Later in the search these participants tended to show more care in scanning the images, enabling them to find the target object. This, however, led to a significant increase in search time.

**Websites.** Mean values of performance time and oculomotor measures, as well as rates of ‘miss’ and ‘false alarm’ errors are shown in Table 4.

As expected, total task performance time on traditional sites was higher, as information density on the screen was considerably higher than on flat screenshots. On traditional sites there were 110 clickable and 64 unclickable screen areas (total: 174), while on flat sites there were 78 clickable and 54 unclickable screen areas (total: 132). For this reason, the mean performance time for a single screen area was calculated (for both traditional and flat sites). The results demonstrated that the average processing time for a screen area (including making a decision on the objects’ clickability and clicking the clickable objects) was significantly higher for flat websites.

An analysis of ‘miss’ and ‘false alarm’ error types revealed a significant difference between traditional and flat sites: errors of both types were significantly more frequent on flat sites. It is noteworthy that the percentage of false alarms on flat sites in our experiment (28 %) almost exactly corresponds with the figure for false alarms (29 %) in the research conducted by Idler [8].

**Table 4.** Results for the clickable objects search

Measure	Traditional series Mean ( $\sigma$ )	Flat series Mean ( $\sigma$ )	Student <i>t</i> -value	<i>p</i>
Performance time (sec)	28.0 (5.5)	24.2 (6.7)	4.081	0.001
Time per screen area (sec)	1.45 (0.28)	1.65 (0.46)	-3.622	0.002
Errors: misses (%)	26.0 (9.6)	35.8 (13.3)	-5.498	0.000
Errors: false alarms (%)	16.6 (9.4)	28.0 (16.1)	-4.688	0.000
Fixation duration (msec)	342 (43.6)	351 (44.0)	-0.915	0.373
Saccadic amplitude (deg)	3.63 (0.42)	3.91 (0.48)	-3.282	0.004
Saccadic peak velocity (deg/sec)	141 (11.6)	146 (16.0)	-2.646	0.017

It should be noted that in the web search task oculomotor effects were revealed, which were the reverse of those found in the text and icon searches: a search for clickable objects on the page with flat design was characterized by a higher saccadic amplitude and saccadic peak velocity. However, we are not inclined to interpret these results as evidence in favour of a higher cognitive load when working with traditional sites. In our opinion, a key role here is played by the difference in the characteristics of the stimulus material. These effects may be associated with fundamental differences in the design of traditional and flat sites, which force subjects to use different scanning strategies. Thus, in our experiment on sites with traditional design the number and density of graphic objects was higher and interface control tools were more distinct. After initial orientation this allowed the user to develop a systematic search strategy – seen in the combination of longer fixations and shorter saccadic duration. By contrast, the flat design sites initially contained less graphic and text information, which normally facilitate the search for interface control tools. This made subjects repeatedly perform search activity and return to viewing certain areas of web pages several times, shown by a decrease in fixation duration and increase in the amplitude and velocity characteristics of saccades. Thus, the search on flat sites was more “chaotic”, which had a negative impact on time and accuracy parameters of task performance.

## 5 Conclusions

Our study has shown that flat design is inferior to traditional design in most of the aspects we analysed. Text search, where fonts popular in flat design were used, leads to higher cognitive load than search in texts with traditional fonts, although there was no significant difference in the objective measure – search time. A flat icons search is performed almost twice as slowly as a realistic icons search, and is characterized by a higher cognitive load. Analysis and processing of user interface objects on webpages with flat design takes more time than on traditional websites (calculated per screen area), and is

accompanied by a significantly higher error rate; the difference in oculomotor activity reflects the specificity of traditional and flat webpage design.

Our experimental study supports the opinion expressed by many HCI and usability experts that flat design is a harmful tendency in area of user interfaces, and should be replaced by interfaces based on the design principles developed over decades of research and practice of HCI and usability engineering.

The research was supported by a grant from the Russian Foundation for Basic Research (14-06-00371).

## References

1. App, E., Debus, G.: Saccadic velocity and activation: development of a diagnostic tool for assessing energy regulation. *Ergonomics* **41**, 689–697 (1998)
2. Banga, C., Weinhold, J.: Essential Mobile Interaction Design: Perfecting Interface Design in Mobile Apps. Addison-Wesley, Upper Saddle River (2014)
3. Belveal, R.: Where Have All the Affordances Gone? (2013). <http://belveal.net/2013/03/19/where-have-all-the-affordances-gone>
4. Crosby, M.E., Iding, M.K., Chin, D.N.: Visual search and background complexity: does the forest hide the trees? In: Bauer, M., Gmytrasiewicz, P.J., Vassileva, J. (eds.) UM 2001. LNCS (LNAI), vol. 2109, pp. 225–227. Springer, Heidelberg (2001)
5. Di Stasi, L.L., Antolí, A., Cañas, J.J.: Main sequence: an index for detecting mental workload variation in complex tasks. *Appl. Ergon.* **42**, 807–813 (2011)
6. Enders, J.: Flat UI and Forms (2013). <http://alistapart.com/article/flat-ui-and-forms>
7. Hou, K.-C., Ho, C.-H.: A preliminary study on aesthetic of apps icon design. In: 5th International Congress of the International Association of Societies of Design Research 2013 (2013). <http://design-cu.jp/iasdr2013/papers/1811-1b.pdf>
8. Idler, S.: Flat Web Design Is Here to Stay. Usabilla, Amsterdam (2013)
9. Li, C., Shi, H., Huang, J., Chen, L.: Two Typical Symbols in Human-Machine Interactive Interface. *Appl. Mech. Mater.* **635–637**, 1659–1665 (2014)
10. Nielsen, J.: Windows 8 – Disappointing Usability for Both Novice and Power Users (2012). <http://nngroup.com/articles/windows-8-disappointing-usability>
11. Noessel, C.: Your Flat Design Is Convenient for Exactly One of Us (2014). <http://cooper.com/journal/2014/01/your-flat-design-is-convenient-for-exactly-one-of-us>
12. Page, T.: Skeuomorphism or flat design: future directions in mobile device user interface (UI) design education. *Int. J. Mob. Learn. Organisat.* **8**, 130–142 (2014)
13. Pomplun, M., Reingold, E.M., Shen, J.: Investigating the visual span in comparative search: the effects of task difficulty and divided attention. *Cognit.* **81**, B57–B67 (2001)
14. Renshaw, J.A., Finlay, J.E., Tyfa, D., Ward, R.D.: Designing for visual influence: an eye tracking study of the usability of graphical management information. In: INTERACT 2003, pp. 144–151. IOS Press, Amsterdam (2003)
15. Schneidermeier, T., Hertlein, F., Wolff, C.: Changing paradigm – changing experience? In: Marcus, A. (ed.) DUXU 2014, Part I. LNCS, vol. 8517, pp. 371–382. Springer, Heidelberg (2014)
16. Stickel, C., Pohl, H.-M., Milde, J.-T.: Cutting edge design or a beginner’s mistake? – a semiotic inspection of iOS7 icon design changes. In: Marcus, A. (ed.) DUXU 2014, Part II. LNCS, vol. 8518, pp. 358–369. Springer, Heidelberg (2014)
17. Velichkovsky, B.M.: Hierarchy of cognition: the depths and the highs of a framework for memory research. *Memory* **10**, 405–419 (2002)

18. Walker, R.: The ABCs of Website User Experience (2013). <http://adpearance.com/blog/the-abcs-of-user-experience>
19. Zelinsky, G.J., Sheinberg, D.L.: Eye movements during parallel-serial visual search. *J. Exp. Psychol. Hum. Percept. Perform.* **23**, 244–262 (1997)

# How to Organize the Annotation Systems in Human-Computer Environment: Study, Classification and Observations

Anis Kalboussi<sup>1(✉)</sup>, Nizar Omheni<sup>1</sup>, Omar Mazhoud<sup>1</sup>,  
and Ahmed Hadj Kacem<sup>2</sup>

<sup>1</sup> ReDCAD Research Laboratory, Sfax University and Higher Institute of Computer Science and Management, Kairouan University, Kairouan, Tunisia  
`{anis.kalboussi, nizar. omheni, o. mazhoud}@isigk. rnu. tn`  
<sup>2</sup> ReDCAD Research Laboratory and Faculty of Economics and Management, Sfax University, Sfax, Tunisia  
`ahmed. hadjkacem@fsegs. rnu. tn`

**Abstract.** The practice of annotation is a secular and omnipresent activity. We find the annotation in several areas such as learning, semantic web, social networks, digital library, bioinformatics, etc. Thus, since the year 1989 and with the emergence of information technology, several annotation systems have been developed in human-computer environment adapted for various contexts and for various roles. These ubiquitous annotation systems allow users to annotate with digital information several electronic resources such as: web pages, text files, databases, images, videos, etc. Even though this topic has already been partially studied by other researchers, the previous works have left some open issues. It concern essentially the lack of how to organize all the developed annotation systems according to formal criteria in order to facilitate to the users the choice of an annotation system in a well-defined context and according to unified requirements. This problem is mainly due to the fact that annotation systems have only been developed for specific purposes. As a result, there is only a fragmentary picture of these annotation tools in the literature. The aim of this article is to provide a unified and integrated picture of all the annotation systems in human-computer environment. Therefore, we present a classification of sixty annotation tools developed by industry and academia during the last twenty-five years. This organization of annotation tools is built on the basis of five generic criteria. Observations and discussion of open issues conclude this survey.

**Keywords:** Annotation system · Metadata · Annotation · Tag · Human-computer environment · Classification · Survey

## 1 Introduction

Digital annotation presents an activity central in many important tasks, such as studying, indexing and retrieving. It is an important modality of interaction with digital objects (web pages, text files, programs, etc.). Annotations cover a very broad spectrum, because they range from explaining and enriching an information resource with

personal observations to transmitting and sharing ideas and knowledge on a subject [10]. They can be geared not only to the individual's way of working and to a given method of study, but also to a way of doing research [18]. Moreover, they may cover different scopes and have different kinds of annotative context: they can be private, shared, or public, according to the type of intellectual work that is being carried out [39]. Furthermore, annotations are not only a way of explaining and enriching an information resource with personal observations, but also a means of transmitting and sharing ideas to improve collaborative work practices [29]. Thus, the possibility of enriching digital contents by adding annotations has attracted many researchers, who have looked at this opportunity from many different perspectives and with a number of purposes in mind [9].

Therefore, since the year 1989, several annotation systems have been developed in human-computer environment adapted for various contexts and for various roles. The list of annotation tools is growing every year. The researchers have mostly focused on providing more and more sophisticated interfaces with several annotation functionalities. These ubiquitous annotation systems allow users to annotate with digital information several electronic resources such as: web pages, text files, databases, images, videos, etc. The proposed annotation tools have been adopted in a variety of different contexts, such as content enrichment, collaborative and learning applications, and social networks, as well as in various information management systems, such the semantic web, digital libraries, and databases [11, 20].

Even though the annotation system is so common and familiar, it turns out to be especially elusive when it comes to explicitly and formally classifying it, mainly because it is applied in very diverse areas. Indeed, we usually derive what an annotation is from the particular task to which it is applied, rather than investigating the annotation system by itself in order to understand its features and how to use it. To cite a few examples, if we deal with the semantic web, annotation systems are considered as makers of metadata [2]; in the field of digital libraries annotation tools are treated as a means of appending an additional content [25]; when we talk about databases, annotation systems represent both provenance information about the managed data and a way of creating the database itself [1]; finally, in social networks and collaborative tagging, annotation systems are tools of representing of tags or keywords on different kinds of digital content, e.g. photos, videos or bookmarks [9]. This flourishing of different viewpoints about the annotation systems, which is usually considered as an enrichment of the community of annotation and metadata, reveals a lack of a clear comprehension of how to organize all the developed annotation systems of the literature according to formal criteria in order to facilitate to the users the choice of an annotation system in a well-defined context and according to unified requirements. As a result, there is only a fragmentary picture of these annotation tools with its classifications and features.

The aim of the article is to provide a unified and integrated picture of annotation systems in human-computer environment. This panoramic view is based on a classification of sixty (60) annotation systems developed in the literature during the last twenty five years by industry and academia. This organization of annotation tools is built on the basis of five generic criteria: *annotation type* (computational/cognitive); *category of annotation system* (application/plugin/web site); *type of annotative activity*

(manual/semi-automatic/automatic); *annotated resource type* (text/web page/video/audio/image/database/web service/Doc/HTML/PDFs) and *application domain* (semantic web/social networks/digital library/learning/databases/web services/computational linguistics/bioinformatics).

The rest of this article is organized as follows: Sect. 2 presents a general presentation of annotation systems and a classification of these tools according several criteria; Sect. 3 draws some key observations and a discussion of open research issues about the annotations systems. Finally Sect. 4 concludes this survey.

## 2 Annotation Systems

### 2.1 Definition of Annotation Systems

Annotation systems have been developed since the 90<sup>th</sup> to transpose on electronic document the secular practice of annotation. Then, these systems have gradually taken advantage of processing capabilities and communication of modern computers to enrich the practice of electronic annotation [20]. An annotation system or still called annotation tool or “annoteur”, is a system allowing users to annotate various types of electronic resources with different kind of annotations [29]. Many researchers have been interested in the creation of annotation tools to facilitate the annotation practice of the annotator in digital environment. We can find numerous commercial software and research prototypes created to annotate electronic resources (see Fig. 1).



**Fig. 1.** Several annotation systems are developed in different contexts.

### 2.2 Classification of Annotation Systems

Annotation systems have been popular among both researchers and practitioners. In the literature the number of annotation systems does not stop increasing every year. The researchers have mostly focused on providing more and more sophisticated interfaces with several annotation functionalities. There are various areas of continuing research

and implementation on annotation systems. Researchers have taken different approaches to develop and implement these applications. However, this diversification about these approaches, which is usually considered as an enrichment of the community of annotation and metadata, reveals a lack of a clear comprehension of how to organize all the developed annotation systems of the literature according to formal criteria in order to facilitate to the users the choice of an annotation system in a well-defined context and according to unified requirements. As a result, there is only a fragmentary picture of these annotation tools with its classifications and features.

To unify and integrate the picture of annotation systems in digital environment, we present a classification of sixty annotation tools developed in the literature during the last twenty five years by industry and academia. This organization of annotation tools is built on the basis of five criteria (see Table 1). Each annotation system should focus on a particular *type of annotation*. In each annotation type corresponding to a set of appropriate *application areas* in which the annotation system is developed. Afterward, an annotation system in a particular application domain can be one of *three categories*: (*application/website/plugin*). For each category, the annotation system is necessarily based on a process of *annotative activity* (*manual/semi-automatic/automatic*) and annotates a particular digital *resource type*.

**Table 1.** Criteria of annotation systems classification

CRITERIA		SPECIFICATION	
	Annotation type	Cognitive	Computational
<i>Application domain</i>	Digital library		Semantic web
	Learning		Databases
	Social networks		Computational-linguistics
<i>Annotation system category</i>		Application – Website – Plugin	
<i>Annotative activity type</i>		Manual – Semi-automatic – Automatic	
<i>Annotated resource type</i>		Text / PDF/ HTML/ Doc/ Table/ Web service/ Sound / Image/ Video/ Database	

## 2.2.1 First Criterion: Annotation Type (Computational/Cognitive)

- *Computational annotation (metadata):* Annotation can be considered as metadata, that is, additional data which relate to an existing content and clarify its properties and semantics [6]. With this aim, annotations have to conform to some specifications that define the structure, the semantic, the syntax, and even the values that annotations can assume. Computational annotation is used to annotate resources with metadata to facilitate their use by software agents. It is used in the field of information retrieval, summarization, document classification, indexing etc. Computational annotation is applicable to any type of resource: web pages, text files, databases and even for images and videos. The recipients of this kind of annotation are both people and computing devices. On the one hand, metadata can benefit

people because, if they are expressed in a human-readable form and their format and fields are known, they can be read by people and used to obtain useful and well-structured information about an existing content. On the other hand, computational annotations offer computing devices the means for automatically processing the annotated contents. The computational annotations cover essentially the domains of semantic web [2], databases [42], computational linguistics [15], web services [40] and bioinformatics [35].

- *Cognitive annotation (track of reading)*: This is an annotation that is processed and manipulated by human agents. This category of annotation requires a cognitive and intellectual effort to be interpreted. It is visible and distinguishable on the document. The annotation is considered as a track of the activity of reading, which means that the reader is active and productive. It is the result of the investment of the reader in the consulted document. Cognitive annotations are regarded as additional content that relates to an existing content, meaning that they increase the existing content by providing an additional layer of elucidation and explanation. However, this elucidation does not happen, as in the case of annotations as metadata, by means of some kind of constrained or formal description of the semantic of the annotated object [6]. On the contrary, the explanation itself takes the shape of an additional content that can help people understand the annotated content. However, the semantic of the additional content may be no more explicit for a computing device than the semantic of the annotated content. Therefore, the final recipients of this kind of annotation are people; because a cognitive annotation does not make the annotated object more readily processable by a computer than the same object without annotations. The cognitive annotations cover essentially the domains of learning [3], digital library [28] and social networks [10].

### 2.2.2 Second Criterion: Application Domains

As we saw previously, the electronic annotation can be organized in two main categories: computational annotations which cover essentially the domains of Semantic web, Databases, Computational linguistics, Web services and Bioinformatics; cognitive annotation which stretch essentially the domains of Learning, Digital library and Social networks.

A relevant example of the use of computational annotation is provided by the *Semantic web* [33] initiative promoted by the *World Wide Web Consortium (W3C)*, which aims at enhancing human-understandable data, namely Web pages, with computer-understandable data, namely metadata, so that information is given well-defined meaning, better enabling computing devices and people to work in cooperation. The process of adding metadata to Web pages is called semantic annotation, because it involves the decoration of existing data, such as a piece of text whose content is understandable only to people, with semantic metadata that describe that piece of text, so that computers can process it and thus in turn, offer automation, integration, and reuse of data across various applications [1]. The annotations are represented using W3C recommended standard languages as syntax for describing and exchanging metadata like *RDF (Resource Description Framework)* or *OWL (Web Ontology Language)*. The Semantic Web proposes annotating document content using

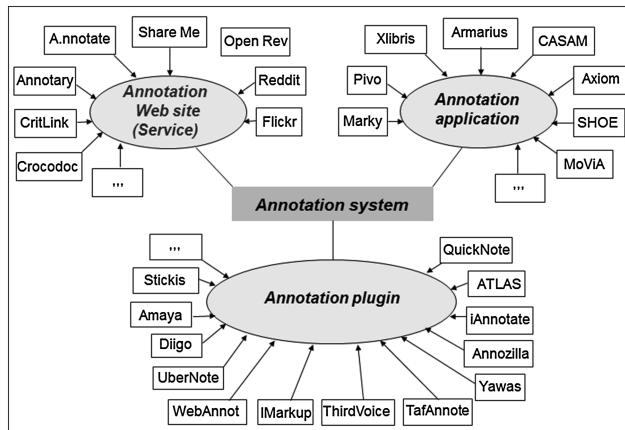
semantic information from domain ontologies. The result is Web pages with machine interpretable mark-up that provide the source material with which agents and Semantic Web services operate.

A broader example of the use of annotations as metadata is provided by the *Computational linguistic* domain, wherein linguistic annotation covers any descriptive or analytic notations applied to raw language data. The basic data may be in the form of time functions, audio, video and/or physiological recordings or it may be textual [15]. The added notations may include transcriptions of all sorts (from phonetic features to discourse structures), part of speech and sense tagging, syntactic analysis, reference annotation, and so on. Corpus annotation, sometimes called tagging, can be broadly conceptualized as the process of enriching a corpus by adding linguistic and other information, inserted by humans or machines (or a combination of them) in service of a theoretical or practical goal.

A relevant example of the use of cognitive annotation is presented in the context of a *Digital library (DL)*. Thus, the creation of new information resources is supported by annotations in two ways. First, when users add annotations to existing information resources, these are new information resources themselves [25]. Second, annotations can also assist in the creation of new information resources. Through annotations, new ideas and concepts can be discussed and the results of such a discussion can then be integrated into the newly created object [1].

Cognitive annotations are also used in the context of *Social networks* wherein social tagging on online portals has become a trend now. It has emerged as one of the best ways of associating metadata with web objects. Social networks have become a popular medium for people to communicate and distribute ideas, content, news and advertisements. Also, social content annotation has naturally emerged as a method of annotating, categorization and filtering of online information [10]. With the increase in the kinds of web objects becoming available, collaborative tagging of such objects is also developing along new dimensions. Social tagging became popular with the launch of sites like Delicious and Flickr [43]. Since then, different social systems have been built that support tagging of a variety of resources. On blogging sites like Blogger, Word-press, Live-journal, blog authors can add tags to their posts. On micro-blogging sites like Twitter [8], hash-tags are used within the tweet text itself. On social networking sites like Facebook, Orkut, etc., users often annotate parts of the photos. Users can also provide tagging information in other forms like marking something as ‘Like’ on Facebook [10]. Social news sites like Digg and Slash-Dot allow users to attach tags to news stories. Yelp, CitySearch and other sites allow users to attach their reviews and other users to select tags to rate reviews too. Multimedia objects like podcasts, live casts, videos and music can also be tagged on sites like YouTube, Imeem, Metacafe, etc. [16].

As a final example of the use of cognitive annotations, we present the context of *Learning*. During this process, the learner’s activities are numerous and especially varied [3, 22]. Thus, the learner can choose to read, write, listen, discuss, experiment, or annotate various resources to achieve his learning goals. Among these activities the annotation practice is very common and omnipresent because while reading, the learner usually uses comments, highlights, circles sections and posts it to annotate the consulted resources [30]. Annotations give students the chance to summarize new ideas



**Fig. 2.** Category of annotation system (Application/Plugin/Website).

while receiving peer support. Tagging can be considered as an action of reflection, where the tagger can summarize a series of thoughts into one or more annotations, each of which stands on its own to describe some aspect of the resources based on the tagger's experiences and beliefs [25].

### 2.2.3 Third Criterion: Category of Annotation System (Application/Plugin/Web Site)

Several annotation systems focus on developing an advanced architecture and building a user-friendly annotation interface to improve annotations of various electronic resources in digital environment. These annotation systems can be classified in three main categories (see Fig. 2):

- **Annotation application:** It is a whole application that allows its users to annotate the resources consulted via the web. Annotation application offers several features and allows multiple browsers to view and annotate web pages. XLibris [26], PDF Annotator [13], InsightNotes [42] and CASAM [14] are examples of annotation application.
- **Annotation plugin:** It is a toolbar or also called additional module added within application or web browser to annotate web pages consulted through the browser. Several toolbars exist such as Amaya [33], Keeppy [18], OntoMat [27], New-WebAnnot [21] and ATLAS [36].
- **Annotation web-site (service):** It is a specialized web-site for annotating that offers to its registered users the ability to annotate web resources. We quote in this type of annotation system: Delicious, Flickr and Reddit [38].

### 2.2.4 Fourth Criterion: Type of the Annotative Activity (Manual/Semi-Automatic/Automatic)

Each annotative activity realized by an annotator using an annotation system passes through two complementary processes: choose the anchor and the shape of the annotation in a given resource (*Process 1*) and specify the properties of the annotation

(Process 2). Based on these two processes, we can classify the existing annotation systems into three categories:

- *Manual annotation system*: Each of the two previously mentioned processes is manually executed by the user himself. The creation process is completely chargeable to the annotator, who selects the shape and the anchor of the annotation and then specifies properties to give to this note. This type of annotation system tries to simply reproduce the annotation process on paper towards computer. Consequently, when it is a question of annotating a wide collection of digital documents, the annotative activity becomes heavy for the annotator. CATool [4] and Keeppy [18] are examples of manual annotation system.
- *Semi-automatic annotation system*: In this type the first process is performed by the annotator while the second process is executed by the annotation tool. The annotator begins to annotate manually. Mean-while the tool textually analyzes his annotations and generates rules of annotation. Then, the tool uses these rules to deduct passages potentially notables and create candidate annotations. The user can then validate or not the annotations proposed by the tool. The system uses these validations to correct its rules. From a certain level of improvement, the semi-automatic annotation system can continue the annotation process alone. This type of annotation system is used especially in the context of the Semantic Web. Indeed, the semantic annotation of Web resources is painful and heavy if it is a question of annotating a large number of resources. Therefore, various tools are proposed to overcome this problem and assist the annotator in the annotation process. WebAnnot [3], Marky [28], New-WebAnnot [22] and Memory-specs [39] are examples of semi-automatic annotation system.
- *Automatic annotation system*: The automatic annotation means that both processes are executed by the annotation system. Thus, according to certain criteria (given by the user or retrieved from another computer system) the annotation tool selects itself the anchor and the shape of the annotation and then specify its relevant properties [23]. One of the most popular automatic annotation system is the search bar which is an extension of Web browsers that allows to highlight with different colors the keywords typed by the user [41]. Share-Me [39], ATLAS [36] and CASAM [14] are examples of automatic annotation system.

### 2.2.5 Fifth Criterion: Type of Annotated Resources

Through the annotations systems, annotators consult and annotate varied electronic resources. Such a resource is dematerialized in a particular format: Text, Doc, HTML, PDF, Table, Video, Image, Audio, Source code, URL, Web service and Database.

Table 2(a, b, c) highlights a set of sixty (60) annotation systems found in the literature review classified according to the five (5) criteria presented above. The tools are presented according to the chronological order their publication or update year (from 1989 until 2014), what allows to evaluate the gradual improvements made to these annotation tools. In the case where the annotation system has several versions, we take the recent version to follow the updates in each annotation system.

**Table 2.** Classification of annotation systems based on five criteria

Name of annotation system	Year	Annotation type	Application domain	Annotated resource type	Category of annotation system	Type of annotative activity
InterNote	1989	✓	Digital Library	Text	✓	✓
Mosaic (version 2.0)	1993	✓	Web	HTML	✓	✓
CommMentor	1994	✓	Digital Library	HTML	✓	✓
Re:mark	1996	✓	Learning	PDF	✓	✓
Third Voice	1999	✓	Web	HTML	✓	✓
AnnotelImage	2000	✓	Bioinformatics	Image	✓	✓
Ann. Sys. for Sem.Web	2002	✓	Semantic Web	Text	✓	✓
NCST	2002	✓	Semantic Web	HTML	✓	✓
IPSA	2003	✓	Digital Library	Image	✓	✓
UCAT	2003	✓	Learning	HTML	✓	✓
AnT&Cow	2005	✓	Semantic Web	HTML	✓	✓
SMORE (version 2.0b)	2006	✓	Semantic Web	Image	✓	✓
WSMO Studio	2007	✓	Web services	Web service	✓	✓
Armarius	2008	✓	Digital Library	Image	✓	✓
Filtered-Push	2009	✓	Digital Library	Database	✓	✓
@Note	2009	✓	Bioinformatics	Text	✓	✓
eduKEN	2010	✓	Semantic Web	Video	✓	✓
Diigo	2011	✓	Social Networks	HTML	✓	✓
Pivo	2011	✓	Semantic Web	Video, Image	✓	✓

(Continued)

**Table 2.** (Continued)

Name of annotation system	Year	Annotation type	Application domain	Annotated resource type	Category of annotation system	Type of annotative activity
		Cognitive	Computational			
AlvisAE	2012	✓	Comp Linguistics	Text	✓	✓
CAT	2012	✓	Learning	All types	✓	✓
eHOST	2012	✓	Comp Linguistics	Text	✓	✓
Surfing Notes	2012	✓	Learning	HTML	✓	✓
Wikilayer	2012	✓	Digital Library	HTML	✓	✓
BioAnnote	2013	✓	Bioinformatics	Doc	✓	✓
CATool	2013	✓	Learning	Video	✓	✓
CLAS	2013	✓	Learning	Video	✓	✓
Indicode	2013	✓	Digital Library	Doc, HTML	✓	✓
In Situ	2013	✓	Semantic Web	Text	✓	✓
Keepy (ver 2.4.6)	2013	✓	Web pages	HTML	✓	✓
MeDetect	2013	✓	Bioinformatics	HTML	✓	✓
Share Me	2013	✓	Semantic Web	Doc	✓	✓
WebAnnot	2013	✓	Learning	HTML, Doc	✓	✓
ATLAS	2014	✓	Semantic Web	Video	✓	✓
Axiom	2014	✓	Digital Library	Doc, PDF	✓	✓
CASAM	2014	✓	Semantic Web	Video	✓	✓
Crocodoc	2014	✓	Learning	Doc, PDF	✓	✓
Domeo Annotation	2014	✓	Bioinformatics	HTML	✓	✓
ndNotes	2014	✓	Digital Library	Doc, PDF	✓	✓

(Continued)

**Table 2.** (Continued)

Name of annotation system	Year	Annotation type	Application domain	Annotated resource type	Category of annotation system	Type of annotative activity
		Cooperative				
		Computational				
Facebook (ver 2014)	2014	✓	Social Networks	All types	✓	✓
Flickr (ver 2014)	2014	✓	Social Networks	Image, Video	✓	✓
Framework ASVA	2014	✓	Semantic Web	Video	✓	✓
Ink Annotation Framework	2014	✓	Digital Library	Doc	✓	✓
InsightNotes	2014	✓	Databases	Database	✓	✓
Marky (ver 2.3)	2014	✓	Bioinformatics	All types	✓	✓
Memory specs	2014	✓	Semantic Web	Text	✓	✓
New-WebAnnot	2014	✓	Learning	HTML	✓	✓
Open Rev	2014	✓	Digital Library	PDF	✓	✓
Org-mode (vers 8.2.10)	2014	✓	Semantic Web	Text	✓	✓
Prosthetic highlighter (v 2.5)	2014	✓	Web	HTML	✓	✓
QuickFox-Notes (ver 2.8.5)	2014	✓	Web	HTML	✓	✓
QuickNote (ver 0.7.5)	2014	✓	Web	HTML	✓	✓
rbutr	2014	✓	Semantic Web	URL	✓	✓
Reddit	2014	✓	Semantic Web	URL	✓	✓
Twitter (ver 2014)	2014	✓	Social Networks	Text	✓	✓
UberNote (ver 2.0)	2014	✓	Web pages	HTML	✓	✓
VideoANT (ver 3.0.0)	2014	✓	Semantic Web	Video	✓	✓
Word Microsoft (ver 2014)	2014	✓	All domains	Doc	✓	✓
YouTube (ver 3.0.0)	2014	✓	Semantic Web	Video	✓	✓

### 3 Observations, Limitations and Open Questions

From the study and the organization of the annotation systems in Human-Computer environment, we can synthesize some observations, limitations and open questions presented as follow:

- *Synthesis 1.* Generally speaking, the term annotation refers to a piece of data associated to another piece of data. Annotations are used in a wide-variety of systems such as blogs, social networks, databases or wikis. Annotations can be defined on every identified resource such as documents, data in a file or in a database, images or video. Annotations can be defined at different level of granularity. For example, in document management systems, annotations can be attached from the whole document to the word level. Annotations can be set manually i.e., made by a user, can be semi-automatic i.e., based on suggestions or fully automated. Annotations can be associated to a group of users (experts, novices, etc.) and can be shared within the group or with other groups. A lot of attempts have been made in the research of annotation systems, many problems, both technical and non-technical, still exist to keep these systems from successful and widely adopted. Marshall [25] argued that “*relatively few researchers are dedicated to understanding what the users actually need*”. Azouaou et al. [3] argued that “*there is nowadays no widespread annotation service*”. Gabriel et al. [10] claimed that “*there is no wide spread use of web annotation systems*”. No single approach is available now that supports all of the features that would work with any standard browser [18]. Also, the existing tools have little connection between each other. Some of them require installation; some require users to log in before use. These factors may all become potential burdens for their users. Many researchers identified several limitations in the implementation of current web annotation systems. The most important of these is the lack of a standardized representation for web annotations, which means that current systems are divergent and proprietary, therefore limiting the possibility for third parties to create clients. Furthermore, they discussed some remaining challenges for annotation systems. The first one is users’ privacy. With the current annotation systems’ architecture and implementation, users’ navigational habits as well as the words they like and dislike can be tracked. When annotations are stored in public annotation servers, the shared or private annotations should be protected from the reach of unauthorized users and the owners of web sites should be able to prevent inappropriate annotations to appear on their sites. A best example of this is this problem faced Third Voice [24]. Thus, annotation systems should take more account the security of annotations made by their users. Another issue is interoperability. Current existing annotation systems adopt different strategies to represent annotations, and use different ways to save these annotations. For example, XPointer have been proposed for XML documents and been adopted by Amaya, and Yawas [7] adds the occurrence of the selected text. A detailed discussion appears in [1] where the authors propose new ways to represent the annotation anchor. Their proposition not only applies to XML documents, but also to HTML, PostScript and PDF documents.

- *Synthesis 2.* In another view point, the designers must take into consideration more user centered collaborative design during creation of annotations systems. Annotation can potentially become a bottleneck if it is done by knowledge workers with many demands on their time. Since few organizations have the capacity to employ professional annotators it is crucial to provide knowledge workers with easy to use interfaces that simplify the annotation process and place it in the context of their everyday work. A good approach would be a single point of entry interface, so that the environment in which users annotate documents is integrated with the one in which they create, read, share and edit them. Annotation systems design also needs to facilitate collaboration between users, which is a key facet of knowledge work with experts from different fields contributing to and reusing intelligent documents [4, 14]. Other issues for collaboration include implementing systems to control what to share with whom. For example, in a medical context, doctors might share all information about patients among themselves but only share anonym information with planners [35]. This brings us to issues related to trust, provenance and access rights. An Intranet provides a more controlled environment for tracing the provenance of annotations than the wild Web but access policies are a critical issue to organizations which are invariably concerned with confidentiality issues for client and staff data. As far as cooperation is concerned, almost all of the analyzed systems show that annotations have great potential for supporting and improving interaction among users, and even among computing devices. Therefore, there is a need for modeling and offering different scopes of annotations, for example, private, shared, or public, and managing the access rights of various groups of users.
- *Synthesis 3.* Although highly sophisticated annotation systems exist both conceptually as well as technologically, we still observe that their acceptance is somewhat limited on behalf of the annotator. Studies made in the works of [25, 31, 32] show that many readers prefer to print an electronic document and to annotate it on paper instead of annotating it directly on its electronic format using an annotation system. Therefore, the process of marking a paper document with the tools that we find in our environment, a pen, a highlighter, is most preferred by the reader instead of reading a document on a screen and the mark via a software interface that requires us to use the keyboard, mouse, stylus, etc. Marshall [25] and Omheni et al. [12, 37] study different kinds of people annotating paper texts for a variety of reasons. That's because annotation on paper is a seamless and flexible practice. Annotations on electronic texts have generally been more problematic. On some reading platforms, annotation is clunky, interrupting reading as the reader pulls up menus, makes selections, switches to a keyboard to type in text: the reader's attention is refocused on the user interface rather than on the book's content. Electronic annotation tools may also limit a reader's expressive intent (e.g., forcing a highlight to be continuous when the reader wants to fragment it, or imposing neatness on a reader when she wants to scrawl). Sometimes the electronic annotations are stored in infelicitous ways so they are either gone when the reader returns to the eBook on a different computer or so they are recoverable when the reader believes them to be deleted and loans the book or document to a colleague. Thus, according to Marshall [25] "*the comfort of annotation on paper cannot be reached. But, the advantages of reading*

*and annotation in digital environment should be able to compensate for this lack of comfort”.*

- *Synthesis 4.* The annotative activity in Human-Computer environment can be manual, semi-automatic or automatic. In the cognitive view point annotation systems tend to spend the manual mode to the semi-automatic mode. While in the computational view point annotation tools attempt to spend the semi-automatic mode to the automatic mode. Neither manual nor automated annotation is infallible, and both have advantages. Manual annotation can be performed with relatively little corpus preparation and can be done inexpensively on quite complex phenomena. It is primarily a question of explaining to the annotators the desired annotation scheme and providing the data. But manual work is slow and limited to small results; mounting a large-scale annotation effort that covers tens of thousands of words is a significant undertaking. Automated annotation, on the other hand, requires a considerable investment in corpus preparation and the programming of the automated tagging system, especially if it is first trained on a seed corpus and then applied to a larger one. Its results may be of poor quality. But it is fast, and can produce outputs over very large corpora of all types in very little time [15]. Several efforts have been made towards building scalable, automatic semantic annotation platforms [34]. Most of these systems focus on manual and semi-automatic tooling to improve the productivity of a human annotator rather than on fully automated methods. However, even with machine assistance, annotation of content is a difficult, time consuming and error-prone task. Semantic Annotation faces the challenge to deliver tools capable of full automatic annotation. The question at this moment is if the challenge of semantically annotating pages is solved and we know it still has a long road ahead.
- *Synthesis 5.* In another view point, the aim of the conceptual annotation models is to formalize the main concepts concerning annotations and to define the relationships between annotations and annotated information resources. Therefore, the proposed formal model captures both syntactic and semantic aspects of the annotations. Thus, there are many types of annotation models available in the scientific state of the art and in the already existing end-user applications. We quote in this survey the three main standard frameworks of the annotation: the W3C annotation project AnnoTea [19], the IEEE Learning Object Metadata (LOM) [17] and the Dublin Core [5]. Using standard formats is preferred, wherever possible, as the investment in marking up resources is considerable and standardization builds in future proofing because new tools, services etc., which were not envisaged when the original semantic annotation was performed may be developed. For annotation systems, standards can provide a bridging mechanism that allows heterogeneous resources to be accessed simultaneously and collaborating users and organizations to share annotations. Because of the need for interoperability, identification and access rights, annotation systems should use basic annotation framework to model the annotation. But despite the success of RDF technology in the Semantic Web; the Dublin core in Web technology and IEEE-LOM in the world of learning, few annotation systems adopt these standards to model annotations. This raises a big problem of interoperability and compatibility between these systems since each of them is based on a particular personal annotation model.

- *Synthesis 6.* Although the existing annotation tools offer good user interaction interfaces, project management and quality control abilities are still limited. Therefore, some recent research works focus on these challenges. For example Marky [28] introduces, a new Web based document annotation tool equipped to manage multi-user and iterative projects, and to evaluate annotation quality throughout the project life cycle. At the core, Marky is a Web application based on the open source CakePHP framework. User interface relies on HTML5 and CSS3 technologies. Rangy library assists in browser-independent implementation of common DOM range and selection tasks, and Ajax and JQuery technologies are used to enhance user-system interaction. Marky grants solid management of inter- and intra-annotator work. Most notably, its annotation tracking system supports systematic and on-demand agreement analysis and annotation amendment. Each annotator may work over documents as usual, but all the annotations made are saved by the tracking system and may be further compared. So, the project administrator is able to evaluate annotation consistency among annotators and across rounds of annotation, while annotators are able to reject or amend subsets of annotations made in previous rounds. As a side effect, the tracking system minimizes resource and time consumption.
- *Synthesis 7.* Finally, most of the annotation systems described previously serve their intended purpose quite well however to deal with the overwhelming size of the web, new approaches need to be considered. In any computerized system, resources (such as processing power, secondary storage, main memory, etc.) are always scarce. No matter how much we have available, a system could always make use of more. These systems are very demanding, much more than normal ones since they need to process huge amounts of texts using complex linguistic tools in the least possible time. Physical resources are not the only bottleneck in the whole system. The raw material which the systems utilize is digital resources, downloaded from the internet. Unfortunately, this process is still extremely slow, especially when several pages are requested simultaneously from the same server. In synthesis, we can conclude that such systems are only usable by top of the range computers having a high bandwidth connection to the internet. A possible solution is to exploit new technologies in distributed computing such as cloud computing [8]. The benefits of such an approach are various. Since the system is utilizing resources elsewhere on the internet, there is no need of a powerful machine with huge network bandwidth. The client is just a thin system whereby the user defines the seed data and simply initializes the system which is then executed somewhere remotely in the cloud.

## 4 Conclusions and Future Work

Based on an overview of existing annotation systems, both in research and industry, this article proposes a unified and integrated picture of annotation systems in several areas of technology. This panoramic view is based on a classification of sixty annotation systems developed in the literature during the last twenty five years by industry and academia.

This organization of annotation tools is built on the basis of five criteria: annotation type (computational/cognitive); category of annotation system (application/plugin/web site); type of annotative activity (manual/semi-automatic/automatic); type of annotated resource (text/web page/video/image/database) and application domain (semantic web/social networks/digital library/learning/databases/Web services/computational linguistics/bioinformatics). The presented list of annotation systems is not exhaustive, but it contains the majority of annotation systems encountered in our survey of annotation tools. Even if there are other annotation systems developed in the literature which are not mentioned in this article, it is certain that these systems can be easily integrated into our classification since the categorization technique is based on cross-cutting criteria applicable for any annotation system.

Nevertheless, the outcome of this article has been limited by the inadequate information about the annotation systems that were discussed. Some of the systems are open source; therefore it is possible to study its documentation and code to explore the structure. However, for many of the other systems, it is very difficult, if not impossible, to get to know their strategies of implementation. Therefore, in the next phase, we will aim to reach a thorough understanding of the implementation and structure of the annotation systems. The tools studied above are necessarily based on annotation models to conceptualize their properties in a formal way to be exploited by computer systems. Thus, we are also planning to propose a survey of conceptual annotation models in digital contents. In another perspective, based on the five criteria of annotation systems classification, we will try to propose a service of annotation systems research. This service presents a user interface providing the possibility of looking for an annotation system which meets the requirements of the user according to the criteria of classification.

**Acknowledgments.** We would like to thank the anonymous reviewers for their extensive comments and suggestions that helped us improve this paper.

## References

1. Agosti, M., Ferro, N.: A formal model of annotations of digital content. *ACM Trans. Inform. Syst.* **26**(1), Article 3 (2007)
2. Andrews, P., Zaihrayeu, I., Pane, J.: A classification of semantic annotation systems. *Semant. Web J.* **3**(3), 223–248 (2012)
3. Azouaou, F., Mokeddem, H., Berkani, L., Ouadah, A., Mostefai, B.: WebAnnot: learner's dedicated web-based annotation tool. *Int. J. Technol. Enhanced Learn.* **5**(1), 56–84 (2013)
4. Barret, A., Celevenger, J., Martini, J.: Open Sourcing Harvard University's Collaborative Annotation Tool. Academic Technology Services, Harvard University (2013)
5. Carmichael, P.: Learning how to learn: using the Dublin Core metadata element set to support teachers as researchers. In: Proceedings of the International Conference on Dublin Core and Metadata Applications: Metadata for e-Communities: Supporting Diversity and Convergence, pp. 201–203 (2002)

6. Caussanel, J., Cahier, J.P., Zacklad, M., Charlet, J.: Cognitive interactions in the semantic web. In: International Workshop on the Semantic Web, Workshop at WWW (2002)
7. Denoue, L., Vignollet, L.: An annotation tool for Web browsers and its applications to information retrieval. In: Proceedings of International Conference on Information Retrieval and its Applications, pp. 180–196 (2000)
8. Dingli, A.: Knowledge Annotation: Making Implicit Knowledge Explicit. Intelligent Systems Reference Library (2011)
9. Ferro, N.: Digital annotations: a formal model and its applications. *Inf. Retrieval Ser.* **22**, 113–146 (2008)
10. Gabriel, H.-H., Spiliopoulou, M., Nanopoulos, A.: Summarizing dynamic Social Tagging Systems. *Expert Syst. Appl.* **41**(2), 457–469 (2014)
11. Gilbert, M., Morgan, J., Zachry, M., McDonald, D.: Indicoder: an extensible system for online annotation. In: Proceedings of the Conference on Computer Supported Cooperative Work Companion, pp. 139–142 (2013)
12. Omheni, N., Kalboussi, A., Mazhoud, O., HadjKacem, A.: Modelling learner's personality profile through analysis of annotation digital traces in learning environment. In: Proceedings of the 15th IEEE International Conference on Advanced Learning Technologies (ICALT 2015). IEEE (2015, in press)
13. GRAHL: PDF Annotator manual. GRAHL software design. (2014). <http://www.pdfannotator.com/>
14. Hendley, R., Beale, R., Bowers, C., Georgousopoulos, C., Vassiliou, C., Sergios, P., Moeller, R., Karstens, E., Spiliotopoulos, D.: CASAM: collaborative human-machine annotation of multimedia. *Multimed. Tools Appl.* **70**(2), 1277–1308 (2014)
15. Hovy, E., Lavid, J.: Towards a 'science' of corpus annotation: a new methodological challenge for corpus linguistics. *Int. J. Transl.* **22**(1), 13–36 (2010)
16. Hwang, W.Y., Wang, C.Y., Sharples, M.: A study of multimedia annotation of web-based material. *Comput. Educ.* **48**(4), 680–699 (2007)
17. IEEE-Standards Association: IEEE Draft Standard for Learning Object Metadata - Corrigendum 1: Corrigenda for 1484.12.1 LOM (Learning Object Metadata). IEEE, Piscataway, pp. 1–24 (2010)
18. Kadam, S., Bajpai, S., Yelmar, P.: Annotation: an investigative survey of annotation types and systems. In: Proceedings of the International Conference on Advances in Engineering and Technology, pp. 102–105 (2014)
19. Kahan, J., Koivunen, M.R.: Annotea: an open RDF infrastructure for shared web annotations. In: Proceedings of the International Conference on World Wide Web, pp. 623–632 (2001)
20. Kalboussi, A., Mazhoud, O., Omheni, N., Kacem, A.H.: An approach of assistance of learner's annotative activity through web services. *Aust. J. Intell. Inf. Process. Syst.* **13**(3), 15–22 (2013)
21. Kalboussi, A., Mazhoud, O., Omheni, N., Kacem, A.H.: A new annotation system based on a semantic analysis of learner's annotative activity to invoke web services. *Int. J. Metadata Semant. Ontol.* **9**(4), 350–370 (2014)
22. Kalboussi, A., Omheni, N., Mazhoud, O., Kacem, A.H.: An Interactive Annotation System to Support the Learner with Web Services Assistance. In: Proceedings of the 15th IEEE International Conference on Advanced Learning Technologies (ICALT) (2015, in press)
23. Kharkate, S.K., Janwe, N.J.: Automatic image annotation: a review. *Int. J. Comput. Sci. Appl. (TIJCSA)* **1**(12), 46–53 (2013)

24. Margolis, M., Resnick, P.: ‘Third Voice: Vox Populi Vox Dei?’. *First Monday* **J. 4.10** (1999)
25. Marshall, C.C.: Reading and Writing the Electronic Book. In: Marchionini, G. (ed.) Morgan and Claypool, Chapel Hill (2009)
26. Morgan, N.P., Schilit, B., Golovchinsky, G.: XLibris: the active reading machine. In: Proceedings of the Conference Summary on Human Factors in Computing Systems, pp. 22–23 (1998)
27. Nacer, H., Aissani, D.: Review: Semantic web services: Standards, applications, challenges and solutions. *J. Netw. Comput. Appl.* **44**, 134–151 (2014)
28. Perez, M., Glez-Peña, D., Fdez-Riverola, F., Lourenço, A.: Marky: a tool supporting annotation consistency in multi-user and iterative document annotation projects. *Comput. Methods Programs Biomed.* **118**(2), 242–251 (2014)
29. Kalboussi, A., Mazhoud, O., Kacem, A.H.: Annotative activity as a potential source of web service invocation. In: Proceedings of the 9th International Conference on Web Information Systems and Technologies, pp. 288–292 (2013)
30. Kalboussi, A., Mazhoud, O., Hadj Kacem, A., Omheni, N.: A formal model of learner’s annotations dedicated to web services invocation. In: Proceedings of the 21st International Conference on Computers in Education (ICCE 2013), pp. 166–169 (2013)
31. Omheni, N., Mazhoud, O., Kalboussi, A., HadjKacem, A.: Prediction of human personality traits from annotation activities. In: Proceedings of the 10th International Conference on Web Information Systems and Technologies (WEBIST 2014), pp. 263–269 (2014)
32. Omheni, N., Mazhoud, O., Kalboussi, A., HadjKacem, A.: The annotation: a track of reader’s personality traits on paper. In: Proceedings of the 2014 ACM Southeast Regional Conference (ACMSE 2014), article No.10 (2014)
33. Quint, V., Carcone, L.: Project Amaya W3C. <http://dev.w3.org/Amaya/doc/WX/Annotations.html>. Accessed on October 2014
34. Reeve, L., Hyoil H.: Survey of semantic annotation platforms. In: Proceedings of the ACM Symposium on Applied Computing. ACM (2005)
35. Rinaldi, F.: Semi-automated semantic annotation of the biomedical literature. In: Proceedings of the 13th International Semantic Web Conference (ISWC), pp. 473–476 (2014)
36. Shah, R., Yu, Y., Shaikh, A., Tang, S., Zimmermann, R.: ATLAS: automatic temporal segmentation and annotation of lecture videos based on modelling transition time. In: Proceedings of the ACM International Conference on Multimedia, pp. 209–212 (2014)
37. Omheni, N., Kalboussi, A., Mazhoud, O., Kacem, A.H.: Automatic recognition of personality from digital annotations, In: Proceedings of the 11th International Conference on Web Information Systems and Technologies (WEBIST 2015), Portugal (2015, in press)
38. Singer, P., Flöck, F., Meinhart, C., Zeitfogel, E., Strohmaier, M.: Evolution of reddit: from the front page of the internet to a self-referential community? In: Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion, pp. 517–522 (2014)
39. Tanaka, K., Kunze, K., Iwata, M., Kise, K.: Memory specs: an annotation system on Google Glass using document image retrieval. In: Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication (UbiComp 2014), pp. 267–270 (2014)
40. Tosi, D., Morasca, S.: Supporting the semi-automatic semantic annotation of web services: A systematic literature review. *Inf. Softw. Technol.* **61**, 16–32 (2015)
41. Tsai, C.F., Hung, C.: Automatically annotating images with keywords: A review of image annotation systems. *Recent Pat. Comput. Sci.* **1**(1), 55–68 (2008)

42. Xiaon, D., Eltabakh, M.: InsightNotes: summary-based annotation management in relational database. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 661–672 (2014)
43. Zheng, N., Bao, H.: Flickr group recommendation based on user-generated tags and social relations via topic model. In: Proceedings of the 10th International Symposium on Neural Networks, pp. 514–523 (2013)

# Mini-Orb: A Personal Indoor Climate Preference Feedback Interface

Markus Rittenbruch, Jared Donovan<sup>(✉)</sup>, and Yasuhiro Santo

School of Design, Queensland University of Technology, Brisbane, Australia  
`{m.rittenbruch, j.donovan, y.santo}@qut.edu.au`

**Abstract.** The control of environmental factors in open-office environments, such as lighting and temperature is becoming increasingly automated. This development means that office inhabitants are losing the ability to manually adjust environmental conditions according to their needs. In this paper we describe the design, use and evaluation of MiniOrb, a system that employs ambient and tangible interaction mechanisms to allow inhabitants of office environments to maintain awareness of environmental factors, report on their own subjectively perceived office comfort levels and see how these compare to group average preferences. The system is complemented by a mobile application, which enables users to see and set the same sensor values and preferences, but using a screen-based interface. We give an account of the system's design and outline the results of an in situ trial and user study. Our results show that devices that combine ambient and tangible interaction approaches are well suited to the task of recording indoor climate preferences and afford a rich set of possible interactions that can complement those enabled by more conventional screen-based interfaces.

**Keywords:** Ambient interface · Tangible interaction · Indoor climate · Individual control · Peripheral awareness

## 1 Introduction

Environmental controls in office environments are becoming increasingly automated. Building management systems (BMS) are being used to control lights, blinds, humidity and temperature. While some of these parameters are set dynamically, based on localised sensor input, they can potentially affect larger numbers of inhabitants, in particular in open office environments. Generally, centralised systems like BMSs do not account for individual inhabitants' preferences regarding indoor climate and environmental office conditions.

The long-term aim of our research is to contribute to the design of systems that aid office inhabitants in controlling their localised office environments, as well as support the process of negotiating shared preferences amongst co-located inhabitants. However, effecting change to environmental conditions in shared offices, based on inhabitant preferences, poses a number of significant technical and social integration challenges. It is a 'wicked problem', which raises numerous interrelated research questions and will require a sustained programme of research well beyond the scope of the study reported here.

As first step towards addressing these challenges this paper focuses on exploring some of the possible *personal input, feedback and interaction mechanisms* such systems might use. We present results of the design, implementation and initial evaluation of a system that uses a range of different ambient and tangible input and output modalities to record subjective office comfort feedback and display sensed data and aggregated group preferences for a range of environmental factors.

Our system consists of three components: (a) a local **Sensor Platform** situated on the users' desks that locally measures temperature, humidity, light levels and noise levels; (b) an ambient and tangible interaction device, called **MiniOrb**, that displays the locally sensed environmental conditions and allows users to select and submit their preference ratings; and (c) a mobile application, **MobiOrb**, that provides an alternative display of the sensed information and input of user-preferences as precise measurements.

Our research aims to address two pertinent questions. First, how can ambient and tangible interaction mechanisms support office inhabitants to record and reflect on their subjective office comfort levels, and second how do these mechanisms compare to screen-based interactions that enable users to see and set the same information with numerical accuracy?

Additionally, our in situ, group-based and real-time preference gathering approach contributes to the range of methods currently available for the study of indoor climate, which is a field still largely reliant on individual and intermittent questionnaire survey methods for gathering personal indoor climate preference data.

## 2 Background

### 2.1 Ubiquitous Computing and Indoor Climate

The field of indoor climate studies emerged in the 1920 s to study people's physiological response to and perception of indoor climate conditions to identify an ideal 'comfort zone' for building climate regulation. This research has focussed on measurable physical parameters such as temperature, lighting levels, sound levels and humidity and has been enormously influential in the development of building standards and legislation for mandated comfort levels [1]. Recently, this idealised and de-contextualized view of what constitutes comfort has begun to be challenged by researchers who emphasize that measureable parameters alone are not enough to give a full picture of the reasons that people perceive the indoor climate the ways they do or account for the actual energy use of buildings [2]. As the introduction to a recent special issue puts it, there is a movement "...away from a passive and toward an active model of the person; away from purely physical or physiological paradigms toward those which emphasize meanings and social settings, and away from universalizing codes and standards ... toward more flexible and more explicitly 'adaptive' strategies in engineering and design" [2, p. 307].

This 'user-centred' turn in building studies [3] highlights questions around how social relations, lived experience, and people's actual use of buildings play into the experience of indoor climate. There is now recognition that achieving energy efficiency in a building is not an engineering problem alone, but a complex and 'wicked' problem dependant on social-relations and patterns of use between inhabitants in a building [1].

There is also recognition of a need to move away from static, pre-defined and steady state models of comfort in order to achieve more sustainable levels of energy use in buildings [2].

Buildings are also increasingly utilising ubiquitous sensing technologies to control the functioning of indoor climate systems in ‘smart’ ways [4]. This often translates into increased automation of indoor climate systems, however it has also been shown that building occupants’ satisfaction levels are strongly negatively affected by lack of control over the environment [5]. Allowing people to control the indoor environment not only improves their overall satisfaction [5] but also can be an effective way to reduce energy consumption [6].

While user engagement in this context can be achieved through a range of interaction techniques we specially consider ambient and tangible interaction mechanisms in the context of this paper.

## 2.2 Ambient Interaction

Ambient devices are a class of interaction mechanism, commonly used to unobtrusively relay information to users. For instance, Ishii [7] explored how to instrument office environments with an array of ambient feedback mechanisms, including lights, sounds, air flow and projected information as part of the ambientROOM environment. Ambient feedback devices have been applied and studied in a wide range of settings [e.g. 8, 9].

Ambient devices commonly rely on relatively simple output mechanisms like LED-based glowing orbs. However, despite their apparent simplicity designers have to carefully consider what information the device should display, how to implement appropriate notification intensity levels and how the device should transition between different states [10]. In addition to the role of ambient devices as pure output mechanisms there is an increasing trend to combine these devices with tangible and other interaction mechanisms in order to enhance people’s physical work environment and provide both input and output capabilities [e.g. 11, 12]. For instance, AuraOrb [13] enhanced a “glowing orb” display with an eye contact sensor and touch input, allowing users to trigger interactions by shifting their focus to the device. Further examples for this approach can be found in the context of instant messaging and presence awareness [e.g. 12, 14, 15].

## 2.3 Informal Awareness

*Informal awareness* addresses tools and mechanisms that facilitate the background awareness between work colleagues, incorporating knowledge of presence, activity and availability [16]. Initial research focused on facilitating casual interaction with the aim to support ongoing collaboration. However, more recent research has explored the notion of informal awareness in the context of domestic environments and other non-work environments [e.g. 17]. For instance Elliot, Neustaedter and Greenberg [18] investigated the contextual properties of location for awareness in the home, showing that where and when devices are deployed is a vital factor for their usefulness and uptake.

In the context of our study suitable *modes of interaction* through which individuals can engage with indoor climate data, need to be accompanied by the consideration of

the *social quality* of these interactions. While the perception of indoor climate is based on individual preferences, the management of shared office environments is an inherently social problem which requires mutual awareness and consensus building across individuals and their specific preferences. As a result we consider aspects of informal awareness as part of our design process.

### 3 Design Process

The starting point for our design was a pre-existing prototype for an embedded wireless sensing platform, which one of the authors had developed in a separate ongoing research project [19]. This platform had been developed to monitor and log indoor climate parameters in an office environment and although it had been programmed to run autonomously the platform did include the possibility for simple user input in the form of a small ‘joystick’ button. Triggered by this, we began to discuss whether it would be possible to expand the possible interactions and feedback available from the platform with a view to collecting user-preference information alongside raw sensor data and as a way of enquiring into future possibilities for collective user control of indoor climate systems.

Given the embedded, wireless and self-contained form-factor of this pre-existing prototype we decided to explore what *ambient and tangible interaction* techniques could be built around the sensor platform. We decided to work with the constraints of the existing platform and that the device should run off the microprocessor of the sensor platform in order that it could be small and unobtrusive enough to be easily positioned on people’s desks. At the same time, we also aimed to make the feedback and interaction of the device rich enough that it would be engaging and useable, so users would actually want to contribute their preference data. We had several overall goals for the design:

- The interactions should be quick and unobtrusive
- The device should provide an ambient awareness of a range of sensor readings
- The device should allow setting of individual preferences in relation to each sensor reading
- The device should allow comparison between individual and group (average) preferences, allowing users to maintain informal awareness of others’ preferences
- The device should allow for user feedback on their level of social connectedness.

In addition to the directly sensed values provided by the sensor platform, we introduced a soft measure of “social connectedness”. We deliberately left the meaning of this measure of “social connectedness” open to interpretation by participants rather than specifying it as for example the number of other people currently in the office. Our aim was to allow people to indicate their feeling of the general social atmosphere of the office *as they interpreted it* and it was these interpretations of participants that we were particularly interested in. Rather than actually quantifying social connectedness, our aim with this measure was to open this up for discussion in our subsequent interviews with participants alongside the other environmental factors (this is discussed further in the “study design” section below).

It is important to note that the design of the ambient interaction device was subject to limitations imposed by the existing sensor platform. The platform had a limited number of input/output ports available that could be used to communicate with the interaction device. As a result the focus of the device design was not to build an interaction device with a large number of possible interaction capabilities. Instead we focussed on how the device could provide a small but sufficient set of interaction mechanisms that would meet our design goals, yet allowed us to use the existing sensor platform infrastructure.

To develop the design we undertook an iterative development process, where we built working prototypes and then ‘lived with them’ ourselves in order to refine the usability, functionality and physical form. Importantly, the programmed ‘behaviour’ of the devices was something that could only be understood by spending some time to experience how it was to interact with the devices over a period of time.

Through this process, several key improvements were made. First was the addition of audio output to provide feedback when setting preferences and to give users a reminder that the device hadn’t been interacted with on any given day. We also discovered that there was a need to support a user in comparing between the current sensor reading and their setting, and to be able to ‘scroll’ through different sensor readings.

We also realised during this process that there was an important question around whether people would want to get a precise reading of sensor data in comparison to the more ambient display provided by the device. This prompted us to design and develop a second prototype based on a mobile-optimised web page, which reproduced the basic functions of the device with the ability to see and set specific sensor values. This second interface represents an alternative approach to building an interface onto the sensor platform where the functioning of the sensor platform is exposed through a web interface. It was therefore useful as a point of comparison for the tangible and ambient design.

### 3.1 MiniOrb System

The MiniOrb system consists of three components, a sensor platform, an ambient and tangible interaction device and a mobile application, each of which fulfil a different role. We introduce each component in turn.

**Sensor Platform.** The MiniOrb *sensor platform* is an Arduino-based sensing device that measures temperature, humidity, light and sound levels via an array of digital and analogue sensors (see Fig. 1, background). Each platform communicates wirelessly across a ZigBee mesh network to a dedicated server. The sensor platforms were placed in a relatively fixed position above users’ desks in order to achieve comparability of sensor readings. The platforms run autonomously and users do not interact with them directly.

### 3.2 MiniOrb Interaction Device

The MiniOrb device is an ambient and tangible interaction device that records user’s office comfort preference values, displays both sensor reading from the user’s local sensor platform, as well as average comfort preferences across all users (see Fig. 1,

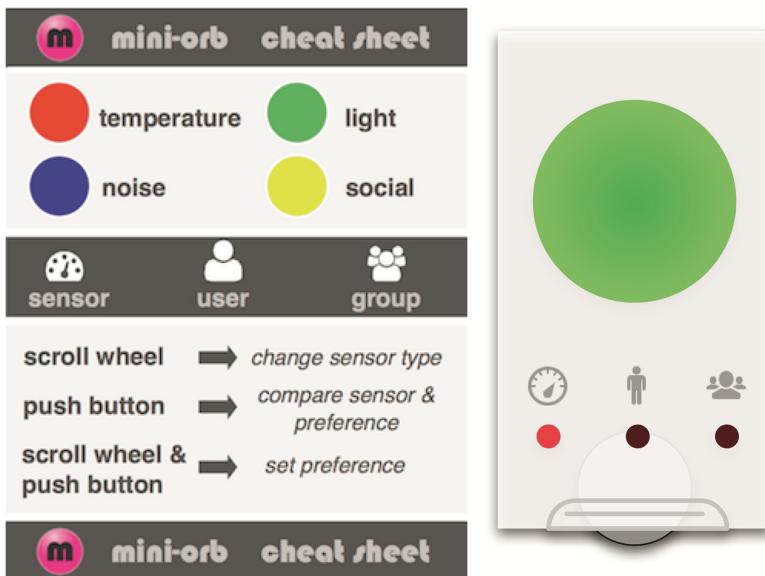


**Fig. 1.** MiniOrb sensor platform (background) and ambient interaction device (foreground)

foreground). The device consists of three small LED's that indicate different states, a piezo speaker, a button and a scroll-wheel potentiometer for user input, as well as a dome-shaped “orb”, a 3D-printed plastic light diffuser which contains a bright RGB-LED and a laser-etched/cut cover.

The dome-shaped “orb” RGB LED is the main output mechanism for the device. To output information the device cycles through a series of colours, which represent different sensor categories of “temperature”, “light”, “noise” and “social”. The available colours were constrained by the capabilities of the RGB LED. The directly sensed parameters of temperature, light and noise were mapped to the primary colours of red, green and blue respectively because these give the clearest colour from the LED, while the ‘soft’ measure of social connectedness was mapped to yellow, which relies on mixing of red and green channels (see Fig. 2, left, for a match between colour and sensor categories). In addition three small LEDs linked to the icons for sensor, user and group respectively indicate whether the readings are a sensor value, a personal preference or a group average. Values are mapped to the colour intensity of the orb, i.e. the higher the value the more intense the colour.

For instance, to display temperature-related information the device cycles through three settings. First it displays the value read by the sensor platform as a matching relative intensity of the colour red. The LED under the “sensor” icon lights up to indicate the state. The device then displays the last known user preference, again indicated by the corresponding status LED “user”. The temperature cycle is then completed by displaying the value for the “group” preference in a similar fashion. Each state is displayed for approximately 5 s. Once the temperature cycle completes, the device moves on to the “light” category using green as the output colour, and so forth.



**Fig. 2.** MiniOrb cheat sheet (left) and conceptual device design (right) (Color figure online)

The “social” category differs from the other categories, in that it is not based on input from the sensor platform, but purely determined by user feedback and intended as a trigger for subsequent discussion of participants’ interpretations of this (we discussed the notion of “social connectedness” in the design section). Thus, for this sensor category the “sensor” value is identical to the “group” value. Each user was given a “cheat sheet” that outlined the colour-codes, states and interactions.

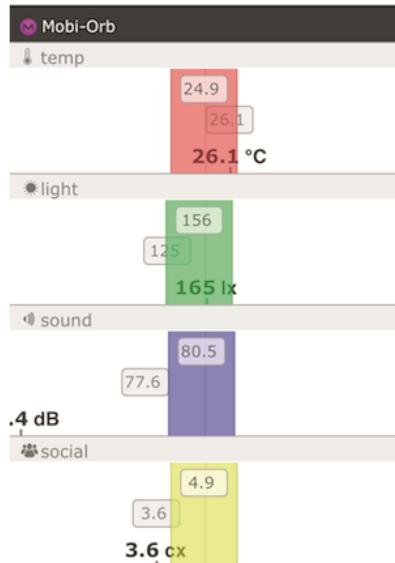
The device offers users three interaction mechanisms by combining the push button and scroll wheel: (1) **scroll wheel**: when users scroll the wheel they can choose one of the four sensor categories manually, e.g. if users are interested in the sound reading they can scroll the wheel to get the device to display the corresponding cycle immediately, without having to wait for the device to complete the other cycles. (2) **push button**: when pressing the button, the device displays the user preference for the current sensor category, and when releasing the button displays the corresponding sensed value. This allows users to efficiently compare their own preference against the sensed value. (3) **scroll wheel & push button**: this function allows users to enter preference values for any sensor category they selected. To do so they keep the button depressed and set the required orb intensity via the scroll. The preference is recorded as soon as the button is released. The device was designed so that this interaction could be easily achieved with a single hand, e.g. by pressing the button with a finger and scrolling the wheel with the thumb.

In addition to the visual output mechanisms the device employs a small number of audio cues to enhance the interaction. The interaction with the scroll wheel is enhanced with subtle “click” sounds that give users a sense of selecting discrete units. A slightly more pronounced sound is used when the wheel moved into the “middle” position.

A separate “chirp” sound is used to notify the user that their preference has been recorded and sent off to the server. Lastly, once per day the device issues a short “remember me” “buzz” sound to encourage users to record their preferences. This sound has been specifically designed to be noticeable, but not to annoy users.

### 3.3 MobiOrb Mobile Application

The MobiOrb mobile application is an alternative interface that provides the same basic functionality as the MiniOrb device, but employs different interaction mechanisms (see Fig. 3). Apart from the interaction approaches, the main difference between the two MiniOrb interfaces is that the mobile interface allows user to interact with specific sensor values (e.g. Temperature 26.1 C).



**Fig. 3.** MobiOrb mobile interface

The main screen consists of four sections, one for each sensor type. Each section contains a colour-coded slider that matches the MiniOrb sensor colour scheme. Users can move these sliders to record their preferences, which are also displayed in plain text in a grey bar in the top part of the slider. The readings in bold at the bottom of each section show the actual sensor value using a sensor-specific unit. The detached grey bar in the middle of each section depicts the group average value. The sensor and preference values exactly match the ones displayed on the MiniOrb device interface. The mobile interface allows users to more accurately assess and set sensor values, but at the same time does not provide the same ambient accessibility as the MiniOrb devices that are situated on users’ desks.

## 4 MiniOrb Evaluation

The evaluation of the MiniOrb system was conducted through a series of user studies. In the context of this paper we report on the outcomes of a two-week in situ trial of the MiniOrb system as well as the outcomes of a series of semi-structured interviews conducted following the trial.

### 4.1 Study Design and Setup

Study participants were recruited amongst the inhabitants of the Queensland University of Technology’s Science and Engineering Centre (SEC), Australia, a multidisciplinary research facility, hosted across two newly constructed buildings, which hosts academics, general staff and postgraduate students. An invitation email was sent out to all SEC inhabitants to participate in the MiniOrb study. The study was structured in three parts: an initial questionnaire exploring existing attitudes towards indoor climate preferences; a two-week trial of the MiniOrb system; and a follow-up interview investigating participants’ experience of use and their interpretation of the sensor categories. Participants could choose how many of the stages to complete based on their own availability. 29 people participated in total in at least one of the stages in the study. 14 of these 29 were available to participate in the initial questionnaire only. 15 participants were available to participate in both the questionnaire and the trials, but again due to availability only 11 of these were also able to participate in the follow-up interviews.

15 participants took part in one of two consecutive trials, which were each conducted over a period of two weeks. At the start of each trial a sensor platform and MiniOrb interaction device were installed on each participant’s desk. Each participant received a short introduction on how to use the device and were given a “cheat sheet” – a very short manual outlining the sensor colour code, symbols and basic functions (see Fig. 2, left). Participants were not instructed to use the device at particular times, but rather encouraged to record preference settings when they felt it was appropriate to do so. The intention of giving the participants the MiniOrb device only in the first week was to allow them to familiarise themselves with what would likely less-familiar interactions and functioning of the ambient and tangible version of the interface than the relatively more familiar mobile phone based MobiOrb interface.

During the second week of the trial participants were offered to use the MobiOrb mobile application in addition to the MiniOrb device on their desk. Our aim was not to compare the two interfaces in an A/B test, but to add an interface that offered accurate numerical readings in order to gain a better understanding of how well the ambient interface performed in relaying office comfort information. Out of fifteen total participants, seven used the mobile interface.

After each trial was completed we conducted a series of semi-structured interviews with the participants, which lasted between 20–30 min. A total of 11 participants across the two trials took part in the interviews. We used a grounded theory approach and conducted open coding to categorise the interview results.

## 4.2 Study Results

In this section we briefly discuss the questionnaire results, but predominantly focus on the results of the trial and follow-up interviews.

**Questionnaire.** The results from the questionnaire provided a baseline for the more detailed qualitative results of the post-trial interviews. The questionnaire results confirmed our assumption that the suggested environmental factors are important to the participants' perception of office comfort. They further confirmed, that with the exception of "social atmosphere" participants felt that they had very limited control over their office environment. Most participants were neither happy nor unhappy with their overall indoor climate, but were reasonably happy with the location of their desks. Generally, our participants felt that being able to change considered environmental factors would have a high impact on their office comfort levels.

**Follow-up Interviews.** The interviews were structured into three overarching sections: (1) attitudes towards office comfort, (2) experience of using the MiniOrb ambient device and (3) experience of using the MobiOrb mobile application. The first section was intended to enhance the data on office comfort levels, collected in the questionnaires, and provided more detail on participants' working context and differences in attitudes between individuals. The other two sections explored both when and how people used the devices on their desk, as well as how they perceived the usability and user experience of the respective device and application. We discuss the results for each section in turn.

*Attitudes Towards Office Comfort.* While many participants appreciated their office environment overall, we identified a number of diverse concerns regarding office comfort. The most commonly mentioned issue was temperature. Many of the participants felt that the overall temperature in the building was set a "little bit" too low. Some participants reported feeling cold at certain parts of the day (e.g. the afternoon). Since the study was conducted in a sub-tropical environment, this generally did not mean that too little energy was used to warm the building, but rather that too much was used to cool it. The second most commonly mentioned issue related to noise. Noise was nearly exclusively interpreted as noise caused by conversations. A number of participants felt disturbed when other people nearby chatted or conducted phone conversations. About half of the participants mentioned that they coped with this interruption by using headphones. Other participants' strategy involved moving to a different (quieter) desk, a meeting room, or working in the library. Participants who reported noise issues were exclusively situated in the open office environment. Other, non conversation-related, background noise was not perceived to be an issue. Lighting, and in particular the setting of the window blinds, was reported as an issue by some participants. Depending on where their desk was located in relation to the windows, they either perceived that they received too much light, which caused issues with glare and reflection on monitors, or the opposite, that the office was too dark and they were not able to see the outside environment. However, complaints regarding lighting were overall less prevalent and intense compared to those regarding noise and temperature. Another issue that was mentioned a number of times was the notion of privacy in the open office setting.

Some participants reported that they would like to have higher, more secluded, cubicles or offices to be able to work in a more private setting. When asked how they perceived their current level of control over their environment, the majority of participants felt that their level of control was very low or even non-existent. The most requested control factor was being able to change the temperature, followed by control over the window blinds. Some participants mentioned that they would like control over aspects like privacy and noise, but also reflected that this would likely require changes in the physical setup of the office.

*MiniOrb Device Experience.* All interviewees reported having used the device. We identified a number of usage patterns with regards to when participants recorded comfort preferences. First, many study participants used the device in the morning when they first arrived at their desk, and again when they returned to the desk from a break. The reported reason for this was that the device was perceived as very inconspicuous (ambient) and participants generally “forgot” that it was there after a while. However, when they returned to their desk they commonly noticed the glowing orb and “remembered” that the device was there. Second, participants would specifically use the device when they became aware of being uncomfortable or when the local environment changed (e.g. the window blinds going up and down). Third, participants commonly entered data when the device issued a “remember me” buzz sound. Nearly all participants perceived this mechanism very positively. They felt that it helped them to remember to provide input and did not feel that the interaction was intrusive or distracting. One participant also reported that they were encouraged by hearing other people send feedback from their own devices (by hearing the “feedback submitted” sound) and subsequently remembered to use the device themselves.

Overwhelmingly, participants enjoyed having the device situated on their desk and perceived that the device was very unobtrusive as well as easy to use. However, specific functionalities were used at different rates as well as interpreted and applied differently. A significant difference emerged in the way people recorded preferences. Some participants used the push button feature, that allowed them directly compare the current sensor reading, for a specific category, with their user preference. These participants would then set the value a “little bit” higher or lower than the current status to indicate gradual preference change. By contrast, other participants would turn their preference value to the maximum or minimum setting to indicate their strong desire for this value to change respectively. These participants did not perceive that they were setting a specific value, but rather interpreted the interaction as “casting a vote”. Participants reported that they particularly engaged in this type of voting when they felt strongly about their choice or wanted to communicate their displeasure (e.g. they felt annoyed because the environment was too noisy to concentrate).

There were a limited number of reported uses of the “social” category. As described earlier, our intention with adding this category to the interface was in large part to trigger discussion in the interviews of what people’s interpretations of it were. While some participants reported that they were unsure how to interpret this category, others gave surprising examples of a use of the feature that resulted in social interaction. For instance some participants belonging to a working group would “turn

up” their social preference value at the end of some working day to mutually indicate to each other that they were ready to engage in social activities. In these cases it seems that the social category functioned not only as a measure of social atmosphere, but also as much as a means for people to signal calls for social action to one-another.

The functionality that was reportedly used least was the “group average” feature. Only some participants reported that they actively observed the group setting after they submitted their preference in order to understand how other users felt. Many other participants however stated that they did not pay attention to the group setting, or in some cases were not sure what it meant.

A number of users pointed out that setting feedback levels made them feel like “somebody cared”. While these participants were aware that the system only recorded feedback values and did not affect change, they nevertheless valued the fact that their opinion did in some way count. One user opined: “*(...) it just gave me the feeling that somebody maybe cares somewhere*”.

The interviews revealed a number of other, smaller, issues regarding the system’s functionality. One participant thought that the “press button” function would allow them to compare personal preference with group average values, rather than sensor values. A single participant felt that the light from the orb was somewhat distracting and subsequently positioned it out of sight. However, this attitude was not shared by the large majority of participants.

*MobiOrb Application Experience.* Out of the eleven participants we interviewed, seven had used the mobile application. The most common observation was that the mobile application was less noticed or thought of. Most participants felt that the ambient device reminded and encouraged them to use it because it was situated on people’s desk. The mobile application, by contrast, had to be remembered and used on purpose.

However, when people actually used the application they appreciated the ease with which feedback values could be set and found it generally easy to use. One participant commented that setting multiple values was quicker and easier on the mobile device. The fact that the mobile device displayed concrete values rather than relative colour hues was an obvious difference between the two interfaces. Our participants on average did not seem to prefer either way of presenting values over the other. Some participants expressed that seeing the concrete values, and in particular the range within which the value could be changed, enhanced their experience: “*It just felt like I knew more what I was saying with the range*”. However, another participant mentioned that he liked being able to focus on setting their perceived comfort levels relative to the current sensed value, without having to think about absolute numbers.

## 5 Discussion

### 5.1 Discussion of Interview Results

The interviews provided a nuanced picture of participants’ attitude towards office comfort and their use of the different elements of the system. In the following discussion we will highlight five pertinent issues that warrant further discussion:

**“Protest” vs. Gradual Vote.** Due to the fact that the feedback mechanism of the ambient device was based on colour intensity, the meaning of feedback values was open to interpretation. Our participants used the feedback mechanisms in two significantly different ways: (a) to submit *gradual changes* based on the sensor value to indicate relative shifts in required comfort levels or (b) to submit a *radical change* by setting the value to the minimum or maximum setting.

The latter approach, here also referred to as a “*protest vote*” was used to express a strong feeling of discomfort and was similar to a yes/no voting approach, while the former approach aimed to provide an accurate reading of the desired value. Both approaches are valid, however the protest vote was less applicable on the mobile application, since users were able to see the specific value of their preference setting. Our participants reported that once they saw the results of their “*protest vote*” on their mobile interface they realised that they had set the preference value either very low or very high and that this setting did not reflect their actual preference. However, we believe that both approaches are valid in the context of providing feedback on comfort levels and should be supported. This issue requires further reflection on the design of future iterations of our system and other similar systems.

**Minimal Design Trade-off.** The minimal design of the interaction device was an important design consideration. This was a design constraint that we consciously chose by deciding to work within the existing constraints of the sensor platform. The challenge was to build a small device that combined suitable ambient output mechanisms with a small number of tangible interaction mechanisms. The device had to support a suitable range of functionality without burdening the user with too much complexity. Based on the results of the interviews we believe that we overall succeeded in achieving this goal, but for future work we would strongly consider redesigning some of the underlying platform while maintaining the philosophy of keeping the platform as minimal as possible.

With regards to its “ambient quality” the device was perceived as fading into the background and being available when people wanted to interact with it. However there were signs that not all of the intended functionality was used to the same extent. In particular, the group average reading was only used by a limited number of participants. This fact is possibly related to our choice of functionality that allowed users to compare the feedback value against the sensor value, but not the group average value. This is a potentially significant design decision because as became clear from the interviews an important factor for people is that they feel that their preferences are reflected or supported by the group. This highlights that indoor comfort is as much a social phenomenon as a measurable physical phenomenon and by choosing to compare with the sensor values rather than the group preference, our interface emphasised the “physical” view.

This presents a design trade-off when dealing with a device with limited interaction mechanisms. We suspect that rather than trying to integrate both the comparison of sensor values and group averages into a single device, an alternative and potentially better design would be to remove rarely used functionality (e.g. group average) and represent this functionality on a separate device or interface (e.g. a “MaxiOrb” with the sole purpose of publicly displaying group averages to a group of users in a section of an office).

**Somebody Cares Somewhere.** The notion that some participants felt positively about the fact that their feedback was recorded highlights the importance of aspects of office comfort that go beyond measurable factors, such as “being appreciated”. With regards to the design of similar systems, this raises the question how systems can be designed to more actively give user the feeling of being listened to as well as finding mechanisms to affect change or reflect office comfort attitudes to other inhabitants (e.g. a “MaxiOrb” public display, mentioned above could indicate that several users felt that the workspace was getting too noisy, and thus raise the level of awareness regarding shared attitudes in office environments).

**Prompting Interaction.** The small “remember me” buzz sound prompt, issued to encourage users to submit a preference value, had a significant impact on the usage pattern of the device. Interestingly, our participants did not find this interaction to be distracting, but perceived it as a welcome reminder to interact with the system. Conceptually, this interaction can be interpreted as briefly moving the device from its’ ambient state into the user’s focus, acting as a *reverse notification*, requesting user interaction, rather than indicating a change in the systems’ state.

**Ambient vs. Mobile Interaction.** It is too early, and beyond the scope of this paper, to conduct a conclusive comparison between the use of the ambient interaction device and the mobile application in the context of our study. However, the results of our interviews indicate that both interfaces fulfilled different and important roles. One of the most important aspects of the interaction device was its ambient nature. The fact that the device was located on people’s desk meant that it acted as a constant reminder, a central quality when seeking to continuously solicit user input. The mobile device by comparison was appreciated for its straightforward and precise interface, which allowed users to provide specific feedback and understand the range of different sensor categories. Interestingly, a number of users remarked that they would have preferred if this interface was located on their computer desktop rather than their mobile phone to provide better integration with the working environment on their desk. Generally, the mobile interface was perceived as an extension that provided additional functionality to the ambient interaction device, rather than a replacement of it.

## 6 Conclusions

In this paper we described the design, use and evaluation of MiniOrb, a system that employs ambient and tangible interaction mechanisms to allow inhabitants of office environments to report on subjectively perceived office comfort levels. One attraction of a tangible interaction approach in this context is that it gives physical presence to a phenomenon that is normally at the background of peoples’ experience. Our research addresses two pertinent questions. First, how can ambient and tangible interaction mechanisms support office inhabitants to record and reflect on their subjective office comfort levels, and second how do these mechanisms compare to more traditional approaches that enable users to see and set specific sensor values?

The results of our study show that minimal interaction devices, combining ambient and tangible interaction approaches, are well suited to engage users in the process providing preferences. This process can be aided by the provision of alternative interface mechanisms that provide accurate sensor and reference values when required. The results of our study are particularly relevant in light of the fact that our system did not affect change in the users comfort levels, but merely recorded their preferences, thus providing less of an incentive to engage with the system. The fact that our system was used and users felt that they were “listened to” highlights the importance of exploring mechanisms to provide individualised control over office comfort levels. While the introduction of our system was successful, the results of our study revealed many nuances with regards to how people provided feedback, which functionality to integrate in a minimal interaction device, how to prompt interactions and the different ways people interpret vague and specific sensor readings.

An important contribution of our approach for ongoing research into understanding people’s responses to indoor climate conditions is that it provides a method of recording preferences *in situ* and *through time* and for encouraging people to reflect on their experience of indoor climate. This supports the need for moving away from static steady-state approaches to indoor climate control to ones that take account of individual variability and changes over time.

## References

1. Jaffari, S.D., Matthews, B.: From occupying to inhabiting – a change in conceptualising comfort. In: IOP Conference Series: Earth and Environmental Science, vol. 8(1), pp. 1–14 (2009)
2. Shove, E., et al.: Comfort in a lower carbon society. *Build. Res. Inf.* **36**(4), 307–311 (2008)
3. Verbeek, P.-P.: *What Things Do: Philosophical Reflections on Technology, Agency, and Design*. Pennsylvania State University Press, Pennsylvania (2005)
4. Liu, X., Akinci, B.: Requirements and evaluation of standards for integration of sensor data with building information models. In: Computing in Civil Engineering. American Society of Civil Engineers, pp. 95–104 (2009)
5. Frontczak, M., Wargocki, P.: Literature survey on how different factors influence human comfort in indoor environments. *Build. Environ.* **46**(4), 922–937 (2011)
6. Brager, G.S., Paliaga, G., de Dear, R.: Operable windows, personal control, and occupant comfort. *ASHRAE Trans.* **110**(2), 17–35 (2004)
7. Ishii, H., et al.: ambientROOM: integrating ambient media with architectural space. In: CHI 1998 Conference Summary on Human Factors in Computing Systems 1998, pp. 173–174. ACM, Los Angeles (1998)
8. Prinz, W., Gross, T.: Ubiquitous awareness of cooperative activities in a Theatre of Work. *ITG Fachbericht* **168**, 135–144 (2001)
9. Chang, A., et al.: LumiTouch: an emotional communication device. In: CHI 2001 Extended Abstracts on Human Factors in Computing Systems 2001, pp. 313–314. ACM, Seattle (2001)
10. Matthews, T., et al.: A toolkit for managing user attention in peripheral displays. In: Proceedings of the 17th Annual ACM Symposium on User Interface Software and Technology 2004, pp. 247–256. ACM, Santa Fe (2004)

11. Ishii, H., Ullmer, B.: Tangible bits: towards seamless interfaces between people, bits and atoms. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems 1997. ACM Press, Atlanta (1997)
12. Hausen, D., et al.: StaTube: facilitating state management in instant messaging systems. In: Proceedings of the Sixth International Conference on Tangible, Embedded and Embodied Interaction 2012, pp. 283–290. ACM, Kingston
13. Altosaar, M., et al.: AuraOrb: social notification appliance. In: CHI 2006 Extended Abstracts on Human Factors in Computing Systems 2006, pp. 381–386. ACM, Montreal (2006)
14. Peek, N., Pitman, D., The, R.: Hangsters: tangible peripheral interactive avatars for instant messaging. In: Proceedings of the 3rd International Conference on Tangible and Embedded Interaction 2009, pp. 25–26. ACM, Cambridge (2009)
15. Greenberg, S., Kuzuoka, H.: Using digital but physical surrogates to mediate awareness, communication and privacy in media spaces. Pers. Technol. 3(4), 182–198 (1999)
16. Rittenbruch, M., McEwan, G.: An historical reflection of awareness in collaboration. In: Markopoulos, P., De Ruyter, B., Mackay, W. (eds.) Awareness Systems: Advances in Theory, Methodology and Design, pp. 3–48. Springer, Heidelberg (2009)
17. Greenberg, S., Neustaedter, C., Elliot, K.: Awareness in the home: the nuances of relationship, domestic coordination and communication. In: Markopoulos, P., De Ruyter, B., Mackay, W. (eds.) Awareness Systems: Advances in Theory, Methodology and Design, pp. 73–96. Springer, New York (2009)
18. Elliot, K., Neustaedter, C., Greenberg, S.: Time, ownership and awareness: the value of contextual locations in the home. In: Beigl, M., Intille, S.S., Rekimoto, J., Tokuda, H. (eds.) UbiComp 2005. LNCS, vol. 3660, pp. 251–268. Springer, Heidelberg (2005)
19. Santo, Y., Loh, S., Fernando, R.: Open up the building : architectural relevance of building-users and their participations. In: Stouffs, R., et al. (eds.) 18th International Conference of the Association of Computer-Aided Architectural Design Research in Asia (CAADRIA 2013), pp. 385–394. National University of Singapore, Singapore (2013)

# Prototyping the Self-Authored Video Interview: Challenges and Opportunities

Stephen Snow<sup>(✉)</sup>, Markus Rittenbruch, and Margot Brereton

Queensland University of Technology, 2 George Street, Brisbane, 4000, Australia  
`{steve.snow,m.rittenbruch,m.brereton}@qut.edu.au`

**Abstract.** Self-authored video- where participants are in control of the creation of their own footage- is a means of creating innovative design material and including all members of a family in design activities. This paper describes our adaptation to this process called Self Authored Video Interviews (SAVIs) that we created and prototyped to better understand how families engage with situated technology in the home. We find the methodology produces unique insights into family dynamics in the home, uncovering assumptions and tensions unlikely to be discovered using more conventional methods. The paper outlines a number of challenges and opportunities associated with the methodology, specifically, maximising the value of the insights gathered by appealing to children to champion the cause, and how to counter perceptions of the lingering presence of researchers.

**Keywords:** Self-Authored video · Electricity · Eco-feedback · Family dynamics

## 1 Introduction

In this paper we discuss our experience prototyping a method we developed for investigating the use of situated technology in the home, called Self Authored Video Interviews (SAVIs). We developed the SAVI process as a means of encouraging *all* members of participating families to become involved in discussions around domestic energy consumption and their in-home energy use displays (eco-feedback systems). Methodologically, SAVIs are situated in between a form of design probe [8], qualitative interview, video diary [6] and the use of video as a design material more generally [2, 13]. It resembles a design probe in that it is a temporary deployment of research technology in the home requiring creativity on the part of the recipient; a temporary imposition or disruption to domestic life. Yet we abstain from calling it a design probe in the truest sense [8], as it is more structured in the nature of interaction it promotes between the participant and the probe, sharing similarities with qualitative interviews. Despite this we found our initial SAVIs to reveal surprising and innovative perspectives and to be potentially useful as a design material. SAVIs also share similarities with video diaries [6] in that they encourage participants to document their personal opinions and feelings on video. They differ, however, in that SAVIs do not require regular updates or diary entries as per [6], whose participants created regular video diary entries with a laptop webcam. In contrast, SAVIs emphasise the participant as a film maker and creative

director, utilising portable video cameras and a request that participants exercise their own artistic licence.

In the following paragraphs we document our experiences developing and prototyping the SAVI method, in particular its fit within the design space of energy consumption in the home. The concern of the paper is chiefly methodological and we do not discuss technical details of the energy consumption feedback system or the nature of its use by the families in great depth; these findings are reported elsewhere [11]. Rather, we document the experiences of our participants with the SAVIs and the characteristics of the methodology that led to the unique insights we gathered into families' energy consumption and the taken-for-granted assumptions made in this sphere. Also, we find that despite self-authored video being advocated as a means of reducing the intrusion of the researcher in the home [13], that some participants are still conscious of the researchers' "presence" as embodied in the camera itself. In light of these experiences, we make some suggestions for design and discuss some challenges and opportunities for how SAVIs might be modified to enhance the utility of participants and the quality of the data. These include: appealing to children to champion the SAVI cause rather than adults and ways of avoiding the perception among participants of the researchers owning the studied technology, or being embodied in (or on the other end of) the video cameras. We conclude by advocating SAVIs to be a useful means of encouraging participation from the whole family in discussions around energy consumption in the home and the design of technologies in this sphere.

## 2 Background

Video has long been identified as a strongly participatory media [9], where: "*participation and emphatic engagement has to be invested to make sense of the material*" [2, p. 1 citing 9] and has an established history of use as a design tool by HCI and participatory design practitioners [2, 3, 13]. In intimate settings such as the home environment, handing control over the creation of the video to the participants themselves has been advocated as a means of removing the intrusion of a researcher with a video camera [13]. Self-authored video provides a unique perspective through the family's eyes into what they deem to be important enough to video-tape, as opposed to what the researchers deem to be important, thus making visible the taken-for-granted assumptions of everyday life [7, 13]. This sheds light on the idiosyncratic functioning of families and individuals [7] and provides insights derived from the moment, rather than from introspection.

In relation to energy use in the home, Pink et al. [10] used a video ethnography based approach as a means of exploring energy use in everyday life, and informing design concepts for energy saving interventions. This involved a process of "video tours" where: "*One or more family members guided the ethnographer around their house... showing us how they make (or seek to make) the home 'feel right'*" [10, pp. 25:6]. This was a means of exploring energy use indirectly. Bourgeois et al. [1] utilised a "technology probe" approach to gather information about the desires and needs of users when learning about the energy generated by their rooftop solar systems. The technology probe

pack included video cameras to invite their participants to create their own self-authored videos. Interestingly, they note that none of their participants felt comfortable recording themselves on the cameras [1].

## 2.1 Context and Rationale for SAVIs

The particular context to which we chose to apply our SAVIs was the social context of domestic energy consumption in the home and eco-feedback [5, 12]. The intent was to acquire a broader participation of householders in discussions around their energy use and their eco-feedback systems, rather than hearing only from motivated lead-users. Despite the family-situatedness of eco-feedback, few studies employ a specific attempt to integrate all household members in the research simultaneously. Where this has been done, it has been found that while one family member may be knowledgeable and engaged with their eco-feedback system, other members of the same family may have no interest in it at all [5, 12]. As such, integrating all family members in these studies may paint a more accurate picture [12]. Our earlier work conducting individual interviews also led us to this conclusion; where the occasional chance participation in the interviews by other family members hinted at a far richer picture being available if participation could be extended to the whole family.

## 3 SAVI Deployment Process

The SAVI methodology was trialled with 12 families living in suburban South East Queensland in Australia. These 12 families had all received an eco-feedback system called ‘Ecosphere’ within the previous six months. The majority of the 12 families were married or de facto and all but three had children living at home. Nine self-authored videos were created in total; two families said they had been too busy and one did not feel comfortable recording themselves. Video cameras (Fig. 1) were deployed in the homes for a period of 1–2 weeks before being picked up.



**Fig. 1.** The deployed video camera

Attached to the cameras were a number of laminated cards (Fig. 1), with the questions: (1) Show us your last interaction with the eco-feedback (2) What features you do like- why? (3) What don't you like about it- why? (4) What has been happening recently- i.e. hot weather, visitors, holidays etc. and (5) What have you learned from it, how did you learned it or who did you learn it from? The family member(s) home at the time of deployment was asked to request that all other family members answer the questions on camera individually. This format encouraged families to work together in the creation of video material. On the sixth flip-card was an optional invitation for the whole family to participate in the “Steven Spielberg Challenge”. Here, each family was invited to make a short “mockumentary” about their energy use or eco-feedback. Suggestions for content included “*outline the roles each family members plays in energy use in the household*”, “*Design your own alternative eco-feedback system*” and other similarly playful scenarios. This open-ended and creative exercise complemented the more structured nature of the first five flip cards. Unfortunately only three families completed the Steven Spielberg Challenges with others citing time constraints or unwillingness.

### 3.1 Analysis

Analysis of the SAVI data was a challenging process, on account of the SAVIs yielding data reminiscent in some ways of qualitative interviews from the first 5 flip cards and much more playful, ambiguous and less structured content from the “Steven Spielberg Challenge”. While we wished to analyse responses from the first five questions, we did not wish to impose order on what was at times highly ambiguous and creative data. Analysis initially took the form of preparing “video cards”, as per the initial steps of [3]. Several stills from each family’s videos were printed onto A4 sheets with the researchers adding annotations and quotes that were spoken at the time of the still. The creation of cards in this manner represented a useful means of capturing points of interest in the audio and video data. Separately, all audio from the videos was transcribed. A thematic analysis process was then undertaken utilising both quotes from the audio alone as well as the pre-prepared video cards. During this process many more video cards were created in order to capture moments, facial expressions and nuances of interactions that were representative of the different themes that emerged during the analysis.

## 4 Overview of Findings

While SAVIs were useful in the practical sense of their role as an alternative for group interviews, the depth of the findings extended far beyond that possible with interviews alone. The process provided a unique and rich insight into the families’ conceptualisation of energy, the use of their eco-feedback and the taken-for-granted assumptions made by people in relation to their energy use. This was key to our becoming aware of the elements of control, responsibility, power and play that occur within families in relation to their energy use. These findings are reported elsewhere in detail [11] and in this section we concentrate instead on sharing some of the issues we found concerning the SAVI methodology and participation.

#### 4.1 Taken-for-Granted Assumptions

Ylirisku & Buur [13] note how self-authored video can shed light on taken-for-granted assumptions that would not become evident with interviews alone. We found several examples of this, which provided considerable insight into family dynamics around eco-feedback. The 13 year old daughter of Family 6 for instance unwittingly referred to electricity in the house as belonging to her dad and identified the eco-feedback as something that was there to save it. She was empathetic to her father's energy saving cause, calling out other family members if she considered them to be wasteful. One example of this was demonstrated in the family's Steven Spielberg Challenge, involving the 13 year old daughter pushing her similar-aged relative into the swimming pool as retribution for getting her in trouble over the air conditioning:



**Girl 1:** “*For god’s sake! You’ve been in bed with the air conditioning on and dad’s blamed me! You’re horrible!*” **Girl 2:** [screams].”

The two teenage boys from Family 9 made a short film for their Steven Spielberg Challenge. This film parodied the human power relations of saving energy in the home and the lead-user of the eco-feedback getting a little too obsessed with his “Spherey”—a nickname for the Ecosphere eco-feedback system. This involved one son dressed up in an apron and spectacles, menacing his co-inhabitants with a spatula.



**Son:** [looks at the eco-feedback, turns to mother] *You were using the heat lamp!*

**Mother:** *I might have been...*

**Son:** *I don't care. That's 10 lashings for you! You know what? I've had it! Me and Spherey need some time alone!*

What is interesting here is that in the pre-install interview, their mother spoke of the boys generally being very good with electricity use and of her not needing to police it. Furthermore, there was at present no father figure in the household, the boys living with their mother who had separated from their father. Thus it appeared the boys were acting out a (humorously dramatised) representation not of what happened in their home, but of what they thought might have been the case in other homes.

## 4.2 Factors Affecting Data Quality

In relation to the quality of data produced by the SAVIs, we found the most informative and comprehensive content came from instances where (1) one family member championed the cause of the SAVI and (2) when the family did not see the camera as an embodiment of us as researchers.

**The Champions:** The 13 year old daughter of Family 6 made the camera her own and took it upon herself to interview not only everyone in her family, but some of her relatives who were visiting from England. This sometimes involved protests from her subjects if they were caught unawares or put on the spot. The championing of the SAVI activity by this girl provided excellent data on the range of opinions and conceptualisations of the eco-feedback by different family members that we may not have been privy to without her insistence. In Family 4, the wife championed the SAVIs and interviewed her somewhat reluctant husband and teenage children, probing them further if she received an inadequate answer. This led to an insight into the family dynamics in relation to the eco-feedback, where it transpired the wife was in fact the only one who ever used it; her husband admitting: *"It just looks a good bit of kit. I don't really know how to use it"* (Husband, Family 4).

**The Lingering Presence of the Researcher:** A key distinction we found in the data was the degree to which different creators of SAVI material conceptualised the camera as an embodiment of us as researchers. The children from Family 6 and Family 9 displayed no intention to cater for the audience of their videos and appeared to derive a lot of fun from creating imaginative, humorous and (for us) very insightful material. Some participants from other families however, particularly adults, appeared much more conscious of the fact that they were providing material that would be watched by us researchers. This was particularly so in Family 2 and Family 7, where some responses appeared cautious and measured. For example, both the wife and adult daughter of Family 7 both began by formally introducing themselves: *"Hello my name is [S] and I am the daughter of [D] and [L] from [house number and street address]"*- Daughter, 21, Family 7. The careful wording and staccato delivery of a response by a 10 year old daughter from Family 2 caused us to wonder whether she may have been instructed by her parents to make sure she said what they thought was appropriate to say, rather than what she thought herself: *"We have learned from the eco-system that rather than using our air con, we have to use the fans and that if we're not using the lights, we have to*

*make sure that they're turned off*- Daughter, 10 Family 2. While this was interesting in its own right, it was the responses from participants who were not self-conscious or concerned about their potential audience which provided the most insight into energy consumption and eco-feedback in the home.

## 5 Opportunities and Challenges for Future SAVI Deployments

Above we have highlighted the creative and insightful data obtained by encouraging participation from the *whole* household in matters of domestic energy consumption through the use of SAVIs. The more structured question cards provided data reminiscent of a qualitative interview, whereas the Steven Spielberg Challenge provided more creative, unexpected and insightful glimpses into family relations around domestic energy use. Based on these examples above, we conclude with reflections on how SAVI's may be improved in subsequent iterations as well as future opportunities for the methodology.

One of the advantages of self-authored video is the ability to gather insights into the function of a family home without the intrusion of a researcher holding a camera [13]. Yet we still found some participants speaking to the camera as if we were in fact still in their home- behind the camera. One challenge we see for future SAVIs is how to transcend this notion. Although we did not design the Ecosphere ourselves, we wonder whether some participants may have thought we did, on account of us conducting the pre-install interviews at the same time as the installers were conducting their own pre-install assessments. We believe that the researchers' disconnection to the device in question is important and that future deployments of SAVIs should be carried out by researchers who are seen to be completely disconnected to the design and install of the technology in question. This might include a different researcher deploying the SAVIs who introduces themselves as independent from the technology.

Another possibility for this is a more specific integration of children in the SAVI's. Despite deploying the SAVIs to adult members of the households, children were vital to the success of our SAVI deployment and were responsible for the creation of some of the most insightful material of the whole exercise. We believe one means of improving the SAVI process is extending the scope of the creative component and, where possible, specifically assigning responsibility for the SAVIs to children, inviting them to take leading roles both behind and in front of the camera. Putting children in specific control of the exercise in this way, and later reporting the content generated, would require that careful consideration be paid to ethical issues such as parental supervision and informed consent for those depicted in the videos. Despite this, we feel placing children more in charge of SAVI deployments would be useful both in generating more insightful SAVI data and countering the noticeable absence of children from HCI explorations of eco-feedback in the home more generally.

As a final reflection, while we advocate the SAVI methodology of combining structure (interview questions) with creativity (Steven Spielberg Challenge), we appeal to future practitioners of this method to pay attention to the analysis of the resultant data. This was a source of tension for us, deploying what can be interpreted as a form of a design probe, but using it to ask five questions of our participants more reminiscent of

a qualitative interview. We are aware of Gaver and colleagues' [4] lament over the tendency (at the time) for authors to analyse, rationalise, and even produce design requirements from probe results- quite at odds with the original ludic and inspirational intent of the original method. Although we do not identify the SAVI's as probes and feel the attachment of interview style questions to the cameras was very successful, we are aware that as a methodology, SAVIs may still sail rather close to this wind. As such, for future SAVI adaptations, while we feel it is still necessary to impose some sort of order, at least on the interview components, we implore fellow authors not to consider the creative components as "*hard data*" [2] or as evidence on which to base requirements for design. Instead, that attention is paid in the analysis process to attempting to understand the world through the eyes of the different family members and use this as a tool for ideation and design inspiration.

In closing, we advocate SAVIs as a useful means of gathering rich data on how families orient towards energy consumption and eco-feedback and how eco-feedback itself shapes family relations. In this paper we have outlined our particular deployment, some of the challenges we faced and some of the opportunities we see for future deployments. We hope other authors will adapt and improve upon this methodology and share their experiences of doing so.

## References

1. Bourgeois, J., van der Linden, J., Price, B., Kortuem, G.: Technology probes: experiences with home energy feedback. In: Methods for Studying Technology in the Home: Workshop paper, CHI 2013, Paris, France (2013)
2. Buur, J., Binder, T., Brandt, E.: Taking video beyond 'hard data' in user centred design. In: Proceedings of the Participatory Design Conference, New York, pp. 21–29 (2000)
3. Buur, J., Soendergaard, A.: Video card game: an augmented environment for user centred design discussions. In: Proceedings of the Designing Augmented Reality Environments, pp. 63–69. ACM Press (2000)
4. Gaver, W., Boucher, A., Pennington, S., Walker, B.: *Cultural Probes* and the value of uncertainty. *Interactions* **XI**(5), 53–56 (2004)
5. Hargreaves, T., Nye, M., Burgess, J.: Keeping energy visible: Exploring how householders interact with feedback from smart energy monitors in the longer term. *Energy Policy* **52**, 126–134 (2013)
6. Iivari, N., Kinnula, M., Kuure, L., Molin-Juustila, T.: Video diary as a means for data gathering with children- Encountering identities in the making. *Int. J. Hum.-Comput. Stud.* **72**, 507–521 (2014)
7. Jacknis, I.: Margaret mead and gregory Bateson in Bali: Their use of photography and film. *Cult. Anthropol.* **3**, 160–177 (1988)
8. Mattelmäki, T.: Design Probes. University of Art and Design Helsinki. Gummerus Printing, Vaajakoski (2006)
9. McLuhan, M.: Understanding Media: The Extensions of Man. MIT Press, Cambridge (1964)
10. Pink, S., Mackley, K.L., Mitchell, V., Hanratty, M., Escobar-Tello, C., Bhamra, T., Morosanu, R.: Applying the lens of sensory ethnography to sustainable HCI. *ACM Trans. Comput. Hum. Interact.* **20**(4), 1–8 (2013). Article 25
11. Snow, S., Vyas, D., Brereton, M.: When an eco-feedback system joins the family. *J. Pers. Ubiquitous Comput.* Published online 10th February 2015 (2015)

12. Strengers, Y.: Smart Technologies in Everyday Life: Smart Utopia?. Palgrave McMillan, London (2013)
13. Ylirisku, S., Buur, J.: Designing with Video. Focusing the User-Centred Design Process. Springer Press, London (2007)

# An Empirical Study of the Effects of Three Think-Aloud Protocols on Identification of Usability Problems

Anders Bruun and Jan Stage<sup>(✉)</sup>

Department of Computer Science, Aalborg University,  
9220 Aalborg East, Denmark  
[{bruun, jans}@cs.aau.dk](mailto:{bruun,jans}@cs.aau.dk)

**Abstract.** Think-aloud is a de facto standard in user-based usability evaluation to verbalize what a user is experiencing. Despite its qualities, it has been argued that thinking aloud affects the task solving process. This paper reports from an empirical study of the effect of three think-aloud protocols on the identified usability problems. The three protocols were traditional, active listening and coaching. The study involved 43 test subjects distributed on the three think-aloud conditions and a silent control condition in a between-subject design. The results show that the three think-aloud protocols facilitated identification of the double number of usability problems compared to the silent condition, while the problems identified by the three think-aloud protocol were comparable. Our results do not support the common emphasis on the Coaching protocol, while we have seen that the Traditional protocol performs surprisingly well.

**Keywords:** Usability evaluation · Thinking aloud · Verbalization · Think-aloud protocols · Empirical study

## 1 Introduction

Software organizations and developers increasingly emphasize usability as a software quality characteristic [21], and usability evaluation is a key tool to improve usability [17, 21]. User-based usability evaluation is an approach to evaluation that produces two types of results [5, 21]: (1) usability measures and (2) usability problems.

Usability measures include various quantitative assessments measured directly or indirectly during a usability evaluation. The three factors in the classical ISO definition of usability are examples of this [13]: effectiveness, efficiency and satisfaction.

Usability problems are a list of problems that have been experienced by users during a usability evaluation that involves usage of the evaluated software. They are typically expressed in a format that enables developers to enhance the usability of a user interface design. Usability problems are also useful as means to understand a particular usability measure, e.g. why efficiency is low [21].

Think-aloud protocols are widely applied in user-based usability evaluations [5, 17, 21]. This technique was originally introduced in psychology to produce rich data through verbalization of cognitive processes [6]. While there is general agreement

about the strength of thinking aloud, there have also been concerns that it affects the task solving process of the test subjects. Yet the amount of systematic empirical studies of the effect of thinking aloud is limited. A recent study of three different think-aloud protocols characterized the literature on think-aloud protocols as being unclear with respect to the protocols applied. With a focus only on usability measures, i.e. effectiveness, efficiency and satisfaction, they found that different protocols affect the usability measures differently [19]. However, they did not deal with usability problems, and previous studies of this point in various directions.

This paper reports from an empirical study of the effect of different think-aloud protocols on the usability problems that are identified in a user-based usability evaluation. The study also replicates a previous study of the effect on usability measures [19], and compares the results on the two types of measures. In the following section, we present related work. This is followed by a description of the method applied in our empirical study. Next we provide the results and discuss these in relation to previous studies as well as their implications for usability practitioners and researchers. Finally, we provide the conclusion and point out avenues of future work.

## 2 Related Work

This section presents an overview of related literature on think-aloud protocols.

### 2.1 The Traditional Think-Aloud Protocol

The think-aloud protocol was originally described by Ericsson and Simon in 1984 [6] as a technique to study cognitive processes through verbalization. They divided verbalization processes in three levels. Level 1 and 2 are verbalizations that do not distract a subject from the current activity and rely only on information in the subject's short term memory. On these two levels, a participant can be probed or reminded to continue verbalizing, e.g. by saying "Keep talking" or "Um-humm". The assumption is that this probe limits distraction from the current task. Level 3 is verbalizations that distract a subject from the current activity and draw on the subject's long term memory. Ericsson and Simon argued that reliable verbal reports can be produced during task solving provided they do not require participants to carry out additional cognitive processing. They asserted that only verbalizations from level 1 and 2 are valid as rich data to understand cognitive processes, because verbalizations from level 3 distract the subject [6].

The challenge for usability evaluators is that they can see what a user is doing but not why. Thinking aloud has been introduced to overcome this. The idea is that when users are working with an interactive system they express what their intentions are and from that the reasons behind their difficulties can be inferred and compared with what they are actually doing [21].

In a usability evaluation with the *Traditional* think aloud protocol, the test moderator will have minimal interaction with the users and only use the phrase "Keep talking" when the user stops thinking aloud. The test moderator will not embark on any conversation or provide assistance of any kind. Thinking aloud is usually related to laboratory-based testing, but it has also been used in other settings, e.g. [1, 3, 11].

## 2.2 Other Think-Aloud Protocols

There is a variety of other think-aloud protocols. Based on field observations of usability evaluations, Boren and Ramey [2] conducted observations of usability practitioners in software companies. They found several discrepancies between the traditional protocol and the way the practitioners conducted the evaluations and in particular the probes they used. They argue that the traditional protocol with an entirely passive listener is unnatural, and “a speech communication perspective may well interpret silence interspersed with commands to keep talking as a more abrasive form of contact” [2, p. 269]. Instead, they propose a protocol that reflects the way human beings naturally communicate, with a combination of statements by a speaker and some of feedback or acknowledgment from a listener. The feedback uses a back channel from the listener to the speaker, confirming for the speaker that there is actually an “active listener”. Boren and Ramey [2] suggest that the feedback that is most effective is the quiet, affirming “Um-humm” response given at relevant times; and if the test subject goes quiet, the active listener can repeat the last word the speaker said with a questioning intonation.

In usability evaluations with the *Active listening* think-aloud protocol, the test moderator will be an active listener providing acknowledging expressions. The moderator will not ask questions directly or start a conversation, but only use verbal expressions like “Um-humm” or “ahh”. If users stop thinking aloud, the test moderator will repeat the last word expressed by the user.

There are several think-aloud protocols in usability evaluation practice that go beyond the traditional and active listener protocols. In practice, a test moderator will often try actively to get the test subjects to talk about their intentions, thinking, understanding and mental model. This is accomplished by using probes that are much more intrusive. Dumas and Redish [5] present “active intervention” as a technique where the test moderator asks questions in an active manner to get an understanding of the test subjects’ mental model of the system and its functioning. Such an intrusive protocol with extensive conversation has been denoted as coaching [19].

In a usability evaluation with the *Coaching* think-aloud protocol, the test moderator will be empathetic and encourage the test subject, e.g. by stating that “everybody has problems with this part of the system” if test subjects express frustration or feel insecure. The moderator may also express sympathy and give confirming feedback, e.g. “well done, you are doing great” when they have completed. When test subjects come to a complete stop, the moderator may assist by asking questions, e.g. “what options do you have?” or “what do you think happens when you click on that menu item?”

## 2.3 Empirical Studies of Think Aloud Protocols

There is a range of empirical studies of various think-aloud protocols. An early study [25] focused on the coaching protocol, showing that the level 3 verbalizations used in this protocol improved the test subjects’ performance, which made them conclude that level 3 verbalizations in usability evaluation imply a bias.

Another early study [9] examined four user-based usability evaluation methods (logged data, questionnaire, interview, and verbal protocol analysis). They found that the verbal protocol analysis was most efficient, and even using two evaluation methods made no statistically significant improvement over the verbal protocol analysis.

A study by Krahmer and Ummelen [14] compared the traditional protocol [6] with active listening [2]. They found that in the active listening condition, more tasks were completed and the participants were less lost on the website. However, the website was untypical, and their active listening protocol included elements of coaching.

A more recent study by Hertzum et al. [10] compared the traditional and the coaching protocol. They found that with coaching, the subjects spent more time solving tasks, used more commands to navigate and experienced higher mental workload.

A study by Rhenius and Deffner [20] focused on the relation between verbalization and where test subjects were looking. By using eye tracking they showed that where test subjects were looking was directly related to what they were verbalizing, and concluded that verbalizations and short term memory are synchronized.

The related work that assesses think-aloud protocols is listed in Table 1. Apparently, these focus mostly on usability measures. We have only found two empirically based assessments of the usability problems identified with different think-aloud protocols. One study compared the traditional with a retrospective think-aloud protocol. The two protocols revealed comparable sets of usability problems [8]. Another study compared the traditional and the coaching protocol. They found that coaching identified more usability problems related to dialogue, navigation, layout and functionality, but the unique problems of this condition were less severe [26]. This limited focus on usability problem identification confirms the general critique of usability research for facing major challenges [7, 24] and having little relevance to practice [23].

**Table 1.** Overview of the literature that assess think-aloud protocols.

	Description of protocol	Assessment of protocol based on	
		Usability measures	Usability problems
Traditional	[6]	[10, 14, 19]	[8, 15]
Active listening	[2]	[14, 19]	
Coaching	[5]	[10, 18, 19, 25]	[26]

## 2.4 Empirical Studies of Thinking Aloud in Practice

It has been argued that usability evaluation practice needs level 3 verbalizations despite their influence on task solving. Practitioners argue that level 3 gives the most useful data for identifying usability problems in interactive software systems [19].

There are only few empirical studies of think-aloud protocols in practice. Nørgaard and Hornbæk [18] observed and interviewed practitioners on the way they conducted usability evaluations. They found that practitioners often asked hypothetical or leading questions, which would elicit level 3 verbalizations [19]. Another study explored the use of think-aloud protocols by usability practitioners. They found that the think-aloud technique was generally appreciated, but divergent practice was reported [15].

Several researchers have requested additional research in the effects of various think-aloud protocols in usability evaluations, e.g. [2, 14], as it is not clear which protocol usability practitioners should employ in different situations.

### 3 Method

We have conducted an empirical study of the three think-aloud protocols in Table 1. The aim was to study the effect of these think-aloud protocols on the identification of usability problems. Part of the study also replicated a study of usability measures [19].

#### 3.1 Experimental Conditions

Our empirical study was designed to compare the following four conditions:

- *Traditional*: the protocol originally suggested by Ericsson and Simon [6], where the test moderator is only allowed to probe with “Keep talking” when the test subject has been quiet for a while.
- *Active listening*: the protocol suggested by Boren and Ramey [2] where the test administration provide feedback or acknowledgment to the test subject by constantly probing with “Um-humm”.
- *Coaching*: the protocol that is generally used in usability evaluation practice where the test moderator is engaged in a continuous conversation with the test subject.
- *Silent*: the test moderator does not talk at any point during the evaluation sessions, except when introducing test subjects to the experiment and tasks. The test subjects are specifically asked not to think aloud.

The first two protocols, Traditional and Active listening, support level 1 and 2 verbalization, thereby providing what is assumed to be the most reliable data. The coaching protocol supports level 3 verbalizations where the probes are expected to initiate use of the long term memory of the test subjects. The Silent condition is a benchmark.

#### 3.2 System

The system evaluated is a data-dissemination website (dst.dk) that provides publicly available statistics about educational levels, IT knowledge and skills, economy, employment situation etc. in Denmark. It is the same type of website as in [19].

#### 3.3 Participants

**Users.** All participating users were invited through emails distributed to the whole university. In total we recruited 43 participating users divided on four different demographical profiles, including 15 participants from technical and administrative

personnel from different departments, 13 faculty members from Ph.D. students to professors from different departments, and 15 students in technical and non-technical educations. All participants were given a gift with a value corresponding to 20 USD.

We distributed the participants evenly according to their demographic profiles over the four experimental conditions, see Table 2.

**Table 2.** Distribution of participants on conditions and demographic profiles, TAP = Technical and Administrative Personnel n = number of participants.

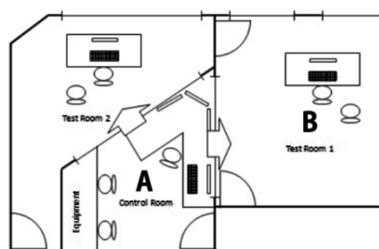
	Coaching (n = 10)	Active listening (n = 11)	Traditional (n = 11)	Silent (n = 11)
Mean age (SD)	34.9 (10.6)	33.2 (10.9)	37.5 (16.5)	40.8 (13)
Females/Males	4/6	5/6	7/4	6/5
TAP/Faculty/Students	5/2/3	3/3/5	4/3/4	4/4/3

**Test Moderators.** We used four test moderators in this experiment, all of whom had previous experience in moderating usability evaluation sessions. We chose to apply a single test moderator for each condition. To avoid test moderators introducing a bias, none of them took part in planning the study, analysing the data or writing this paper.

**Evaluators.** The two authors of this paper analysed all video material from the 43 evaluation sessions. Both have extensive experience in analysing video data. To reduce possible bias from learning, the evaluators analysed the collected data in different (random) orders. The moderator/analyst separation is uncommon in practice, but was necessary to reduce bias from experimenter expectancy.

### 3.4 Setting

All usability evaluations were conducted in a usability lab, see Fig. 1. The test moderator was in the control room (A) and the user in the test room 1 (B). They communicated through microphones and speakers. The rooms were separated by one-way mirrors so the user could not see the test moderator. This physical separation of test subjects and moderators is different from typical practice. It was introduced to remove the influence from the test moderator's body language and other visible expressions.



**Fig. 1.** Layout of the usability laboratory. The test moderator was located in the control room (A) and the user in test room 1 (B).

### 3.5 Procedure

The usability evaluations were conducted over the course of five days. On the first day of the four test moderators each did two evaluation sessions in order to get acquainted with their protocol. They had received written instructions on how to act in their particular protocol three days in advance, to allow for preparation. After completing their two evaluation sessions, they feedback on how they performed with respect to the protocol. The test moderators did not observe each other's sessions.

Each of the following four days was devoted to evaluations with a particular protocol. All were conducted as a between-subjects study and each session lasted one hour. In all evaluation sessions, the users were first asked to complete two tasks using a similar website, to get acquainted with the protocol applied in their session. They were then asked to solve eight tasks using the main web site. At the end of each session, the user filled in a shortened version of the QUIS questionnaire, applied in [19].

**Tasks.** The users solved eight tasks that varied in difficulty. For example, the first task was to find the total number of people living in Denmark, while a more difficult task was to find the number hotels and restaurants with 1 single employee in a particular area of Denmark.

### 3.6 Data Collection

All 43 sessions were video recorded. A questionnaire was collected for each user.

### 3.7 Data Analysis

The analysis of video material was divided in two parts: A joint analysis of 16 videos followed by individual analysis of the remaining 27 videos. All videos were transcribed in a log file before analysing for usability problems, cf. [22].

**Joint Analysis.** In order to achieve consistency in usability problem identification with the different protocols, the authors of this paper first analysed 16 videos together (four videos from each condition). These were analysed in random order.

Due to the variety in protocols we found it important to ensure consistency of problem identification in the individual analysis. To support this, we adapted the conceptual tool for usability problem identification and categorization in [22]. We distinguished used the following ways of detecting a usability problem:

- (A) Slowed down relative to normal work speed
- (B) Inadequate understanding, e.g. does not understand how a specific functionality operates or is activated
- (C) Frustration (expressing aggravation)
- (D) Test moderator intervention
- (E) Error compared to correct approach.

Examples of usability problems are “cannot find the detailed information for a municipality” or “cannot get overview of the presented information”.

**Individual Analysis.** After the joint analysis, we individually analysed the remaining 27 videos in different and random order using the same tool as in the joint analysis. Duplicates where more than one user experienced the same problem were removed (Table 3).

**Table 3.** Mean any-two agreement between evaluators, n = number of datasets analysed individually by the evaluators.

	Coaching (n = 6)	Active listening (n = 7)	Traditional (n = 7)	Silent (n = 7)
Mean (SD)	0.42 (0.11)	0.44 (0.12)	0.47 (0.08)	0.44 (0.17)

**Merging Individual Analysis Results.** After the individual analysis, we merged our two individual problem lists. We did this by discussing each identified problem to determine similarities between the two individual lists. Across the 27 videos we had an average any-two agreement of 0.44 (SD = 0.11), which is relatively high compared to other studies reporting any-two agreements between 0.06 and 0.42 [12]. For the agreement between evaluators, we found no significant differences between the four conditions using a one-way ANOVA test.

**Calculating QUIS Ratings.** The satisfaction score was calculated in the same way as in [19] by combining the sixteen scores from the modified version of the QUIS, each with a score on a Likert scale from 1 to 7. Thus a user could give a total score from 16 to 112 where a high score reflects more user satisfaction with the website.

## 4 Results

In this section we present our findings on usability measures, i.e. effectiveness, efficiency and satisfaction, as well as on usability problems.

### 4.1 Effectiveness

We measured the mean number of tasks completed correctly by users in the four conditions. Users in the Coaching condition had a higher completion rate than users in the other three conditions, while users in the Active listening, Traditional and Silent performed similarly. A one-way ANOVA test reveals no significant differences between the conditions for effectiveness.

We also measured the number of times the users gave up while solving a particular task. Users in the Silent condition had a tendency to give up more often than users in the other conditions, while users in the Coaching condition had the lowest rate in this respect. A one-way ANOVA test reveals no significant differences between any of the conditions with respect to the number of times that users gave up.

## 4.2 Efficiency

Table 4 shows the task completion times in seconds for the four conditions. For each condition, we have calculated the mean value of the total task completion time for each user on all tasks. The users in the Coaching condition on average spent most time while users in the Silent condition had the lowest time. The Active listening and Traditional conditions performed in between these.

**Table 4.** Mean task completion times in seconds, n = number of users.

	Coaching (n = 10)	Active listening (n = 11)	Traditional (n = 11)	Silent (n = 11)
Mean (SD)	238 (42)	219 (47)	220 (50)	175 (72)

A one-way ANOVA test reveals no significant differences between any of the four conditions. However, it should be noted that the difference between the Silent and Coaching conditions is close to significant ( $p = 0.055$ ).

## 4.3 Satisfaction

The QUIS questionnaire included 16 questions where the response for each was a score on a 1–7 Likert scale. Like [19] we added these numbers together, yielding a score between 16 and 112 where a high score should indicate a high level of satisfaction. A one-way ANOVA test of our data reveals no significant differences between median ratings in the four conditions.

The simple sum combines a broad range of different factors. To complement this, we considered a subset of the questions that deal specifically with the users' overall reactions to the website (each on a scale from 1 to 7):

- Terrible – Terrific
- Frustrating – Satisfactory
- Difficult – Easy.

Table 5 shows the overall median scores for these three questions. The Coaching and Traditional conditions gave the highest overall medians, while the Active listening and Silent conditions gave the lowest scores. Note that QUIS measure constructs with more questions than those presented in Table 5. Picking out individual question items would be debatable practice. This is why we considered all three questions of the “overall reactions” subcategory in QUIS, cf. [19].

A one-way ANOVA test reveals significant differences between these one or more conditions ( $F(3,119) = 5.29$ , 5 % level,  $p < 0.002$ ). We apply a Tukey's pairwise comparison test on all pairs of conditions in order to detect which are significant. The Tukey test reveals significant differences in overall median ratings between Coaching–Active listening ( $p < 0.03$ ) as well as between Coaching–Silent ( $p < 0.03$ ). The Tukey test also suggests significant differences in overall median ratings between

**Table 5.** Satisfaction ratings given by users for overall reactions to the website, n = number of users.

	Coaching (n = 10)	Active listening (n = 11)	Traditional (n = 11)	Silent (n = 11)
Terrific-Terrible	4	4	4.5	3
Frustrating-Satisfactory	4.5	3.5	3.5	2
Difficult-Easy	4.5	3.5	4	2.5
<b>Overall Median</b>	<b>4.5</b>	<b>3.5</b>	<b>4</b>	<b>2.5</b>

Traditional-Active listening ( $p < 0.04$ ) and Traditional-Silent ( $p < 0.03$ ). There are no significant differences between Coaching and Traditional, and conversely there are no differences between Active listening and Silent conditions. Thus, users in the Coaching and Traditional conditions express higher satisfaction with the website compared to users in the two other conditions.

The questionnaire also included a question on perceived friendliness of the test moderator that was rated on a 7 point scale from unfriendly to friendly. Users in the Coaching condition gave higher ratings than users in all other conditions, but a one-way ANOVA test reveals no significant differences in these ratings.

#### 4.4 Usability Problems

Through the detailed video analysis, we identified a total of 165 usability problems across all four conditions.

**Number of Identified Problems.** Table 6 shows the mean number of problems identified per user in the four conditions. We identified most problems in Coaching. This is closely followed by Active listening and Traditional, while the Silent condition revealed only around half of the problems identified in the other conditions.

**Table 6.** Mean number of problems identified using the different TA protocols.

	Coaching (n = 10)	Active listening (n = 11)	Traditional (n = 11)	Silent (n = 11)
Mean (SD)	39.7 (13.8)	37.6 (14.6)	36.7 (8.9)	18.7 (6.1)

A one-way ANOVA test reveals highly significant differences between one or more of the conditions ( $F(3,39) = 1.4$ , 5 % level,  $p < 0.001$ ). A Tukey's pairwise test for significance reveals highly significant differences between Silent and Coaching ( $p < 0.001$ ), Silent and Active listening ( $p < 0.002$ ) and Silent and Traditional ( $p < 0.004$ ). We found no significant differences between Coaching, Active listening and Traditional. In other words, the Silent condition revealed significantly fewer usability problems per user compared to Coaching, Active listening and Traditional. This also means that we found no significant difference between any of the think-aloud

protocols. In order to test for type II errors in our statistical analysis, we calculated the  $\beta$  value ( $\beta = 0.02$ ) which led to a statistical power of  $1 - 0.02 = 0.98$ . This is above the threshold of 0.8 [4], which indicates that sample size, reliability of measurement and strength of treatment are acceptable.

**Agreement Between Users Within Each Condition.** Table 7 shows the mean any-two agreement between users in the different conditions. This describes the extent to which users in a condition experienced the same or different usability problems; a high number reflects a large overlap in experienced problems. We found most overlap in the Traditional condition, followed by Coaching and Active listening. The least overlap was found in the Silent condition. Table 8 shows the test statistics comparing each of the conditions with respect to the agreement between user.

**Table 7.** Mean any-two agreement between users, n = number of all pairs of users.

	Coaching (n = 45 pairs)	Active listening (n = 55 pairs)	Traditional (n = 55 pairs)	Silent (n = 55 pairs)
Mean (SD)	0.18 (0.07)	0.16 (0.09)	0.21 (0.08)	0.14 (0.09)

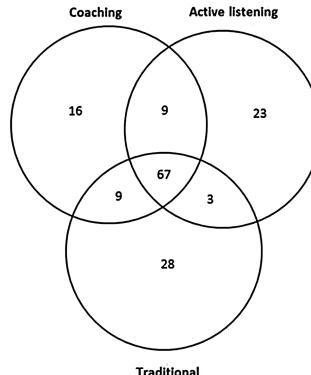
**Table 8.** Pairwise comparisons with respect to any-two agreements between users, □ = almost significant difference, \* = significant difference, \*\* = highly significant difference.

	Coaching	Active listening	Traditional	Silent
Coaching		p>0.8	p>0.4	P<0.02 *
Active listening	p>0.8		p>0.09	P>0.1 □
Traditional	p>0.4	p>0.09		P<0.001 **
Silent	P<0.02 *	P>0.1 □	P<0.001 **	

A one-way ANOVA shows significant differences between one or more of the conditions ( $F(3,206) = 7.32$ , 5 % level,  $p < 0.001$ ). Table 8 shows the results obtained from a Tukeys pairwise comparison test. The Tukey test reveals significant differences between all pairs of conditions and shows significant differences between Silent-Coaching and Silent-Traditional. The differences between Silent-Active listening are close to significant. These findings indicate that users in the Silent condition have a lower overlap in experienced problems compared to the three other protocols that performed similarly.

**Agreement Between Conditions.** The Venn diagram in Fig. 2 shows the overlap between the three TA protocols where 67 problems out of the 155 (43 %) are common between all. Additionally Coaching revealed 16 unique problems, Active listening 23

and Traditional 28. Thus, the Traditional protocol almost reveals twice as many unique problems compared to the Coaching condition. Note that, for simplicity the Silent condition was intentionally left out in Fig. 2. The Silent condition revealed 64 (39 %) of which 54 were identified through TA protocols.



**Fig. 2.** Venn diagram showing overlap in problems between think-aloud protocols.

#### 4.5 Types of Problems Identified

In this section we provide a qualitative account of the types of usability problems identified through each of the protocols. To cover individual differences between the conditions this qualitative analysis is based on the problems uniquely identified by each protocol.

In general we found that the problems descriptions of the Silent condition did not extend beyond the notes made in log files. We would observe a problem when users made an error, e.g. choosing a wrong link. A typical example is “*Wants to find the number of unemployed persons in 2007. Selects the wrong link, but returns to the previous page*”. This is very similar to the entries made in the log files for each user, i.e. they are pure observations, which did not provide enough details to extend problem descriptions with interpretations of the cause of the problems.

In contrast, we found that the Coaching condition in general led to identification of several types of problems. This protocol primarily led to the identification of problems regarding unclear information, e.g. “*Does not understand the meaning of the rows in the table*”. The coaching condition also led to the identification of problems regarding affordance, e.g. “*Wrongfully believes that more details can be obtained by clicking on a table row*”. Finally, Coaching also revealed problems related to visibility such as “*List of tables is immense, difficult to locate the correct one*”.

In case of the Active Listening condition, we found that most problems concerned unclear information. An example of such a problem is “*Difference between menu options is unclear*”. Otherwise there were relatively few problems in other categories such as affordance and visibility.

The Traditional condition primarily led to the identification of problems related to visibility, e.g. “*Does not notice heading of the graph and is therefore looking up numbers in the wrong location*”. Similar to Coaching, the Traditional condition also revealed problems from the other categories concerning unclear information and affordance.

As described above, we found that the three TA conditions led to data from which we could enrich problem descriptions to include notions of why a particular problem was observed. The opposite was the case for the Silent condition where we could observe a problem but not interpret why it occurred. Furthermore we found that the Coaching and Traditional conditions were similar in terms of the breadth of problem types identified. The Active Listening protocol primarily revealed problems related to unclear information.

**Problem Severity.** We categorized all identified problems according to critical, serious and cosmetic [16]. We found that the think-aloud protocols reveal about twice as many critical (mean = 23.7, SD = 2.3), serious (mean = 122.7, SD = 5.6) and cosmetic (mean = 248.7, SD = 8.6) problem instances as the Silent condition. This follows the tendency of the mean number of problems identified per user (see Table 6). There is a comparable distribution of the severity of problems in each of the conditions, e.g. 9 % of problem instances in the Coaching condition are critical, which is comparable to the 7 % in the Silent condition. Thus, we found no significant differences in this respect.

#### 4.6 Categories of Problem Observations

We also categorized the identified usability problems according to the five ways of detecting a usability problem: (A) Problem identified due to time delay, (B) Shortcomings in a user’s understanding of system functionality, (C) User frustration, (D) Test moderator assistance and (E) User makes an error.

As an example, we found that 29 % of all problems identified through the Coaching condition were observed based on time delays (category A). For this category of observations, we found no significant differences between conditions. This was also the case for category B observations.

As another example, we found that 8 % of the problems identified in the Traditional condition were observed on the basis of user frustration (category C). In this case a one-way ANOVA revealed significant differences between conditions ( $F(3,39) = 3.2$ , 5 % level,  $p < 0.04$ ). A Tukey’s pairwise comparison test reveals significant differences between the Silent and Traditional conditions ( $p < 0.03$ ), but no differences between other pairs.

The Coaching condition is characteristic in allowing the test moderator to help test subjects. This is also reflected by the fact that 10 % of the problems identified in the Coaching condition were observed by test moderator assistance (category D).

Finally, it can be seen that 56 % of all problems found in the Silent condition have been observed due to a user making an error (category E). Here we found significant differences between one or more conditions ( $F(3,39) = 15.7$ , 5 % level,  $p < 0.001$ ).

A Tukey's pairwise comparison test shows significant differences between the Coaching condition and all other conditions with  $p < 0.001$  in all cases.

These results show that there are no differences between the four conditions on the proportion of usability problems identified because users were (A) slowed down relative to normal work speed or (B) demonstrated an inadequate understanding of the system. With problems identified because users (C) expressed frustration, significantly more problems were identified in the Traditional condition compared with Silent. For problems identified because of (D) test moderator intervention, there were significantly more problems identified with the Coaching condition compared to all the other. Finally, for problems identified because (E) users made errors compared to correct approach, the Silent condition identified significantly more compared to Coaching.

#### 4.7 Limitations

The empirical results presented above have a number of basic limitations. The study involves three specific think-aloud protocols, and there are many alternatives. We chose the three because they are either classical or commonly used by usability practitioners [5, 19, 21]. Our aim of replicating the previous study made us consider the same protocols as they used, cf. [19].

In each condition, there was one test moderator, and this person was different between the conditions. This was chosen to ensure consistency across all users in each condition. We could have used the same moderator in all conditions, but that could potentially cause confusion and make the conditions overlap. We tried to reduce a possible effect of this by training and assessing the moderators in accordance with the protocol before the tests were conducted.

The empirical study was only based on a single data-dissemination website where users solved 8 pre-defined tasks. To enable replication, we chose the same type of website as the previous study and used exactly the same tasks.

We had 43 users participating in the study. They are not a random or representative sample of the entire population of Denmark. The number of users is limited, but this is comparable to the empirical studies we have in our list of references, where the number of users is between 8 and 40, except for one study based on 80 users [19].

### 5 Discussion

In this section, we compare our results to some of the related work and discuss implications for practitioners.

#### 5.1 Comparison with Related Work

Our empirical study has focused on the consequences of using different think-aloud protocols both on the usability measures and on the usability problems identified.

For usability measures, we have found that users have the highest effectiveness, i.e. they complete more tasks in the coaching condition compared to all other conditions

and that active listening, traditional and silent perform similarly. This is in line with the study presented in [25] where it was found that the coaching protocol improved the test subjects' performance. The study in [19] compared the traditional, active listening and coaching protocols and results from that experiment shows a significantly higher level of effectiveness in the coaching condition than in the other conditions, while there were no significant differences between the traditional and active listening protocols [19]. Thus, we found similar tendencies in favor of the coaching condition in our study but the differences we found were not significant. This is supported in [26] where the traditional and coaching protocols are compared. Furthermore, the study in [14] found a discrepancy compared to our study. In that study the traditional and active listening protocols are compared and it is found that more tasks were completed with active listening. In our study we found almost no difference in effectiveness between these two protocols. However, the result in [14] is not entirely reliable as the active listening protocol included some elements of coaching.

In terms of efficiency we found similar performance between all the think-aloud protocols, while users in the silent condition had considerably lower task completion times. This is similar to [8], who showed that the traditional protocol, due to the requirement of thinking aloud while working, had a negative effect on task performance. Similarly, in [10] the traditional and coaching protocols are compared and findings from that study show that the coaching protocol resulted in higher task completion times. Thus, we do witness similar tendencies as reported in the above literature. However, like [19] we found no significant differences in terms of efficiency.

For the satisfaction ratings, we found that users in the coaching and traditional conditions were significantly more satisfied with the website compared to users in active listening and silent. In a similar study [19] partially agrees they found users in the coaching condition to be most satisfied [19].

For usability problems, we found that the silent condition revealed significantly fewer usability problems than any of the other conditions; and even when we were able to observe problems in that condition, we were often unable to explain why they occurred. This is also supported by the finding that most problems were observed simply by users making errors (category E observations). The study in [8] compared the traditional with a retrospective think-aloud protocol, which revealed comparable sets of usability problems. In contrast, another study has shown that different protocols can reveal different types of problems. The study presented in [26] compared the traditional and coaching protocols and found that the coaching protocol led to a larger number of usability problems related to dialogue, navigation, layout and functionality. Additionally, they found that the problems which were unique to the coaching condition were mainly at a low level of severity [26]. This is partly supported by our findings that the proportion of critical, serious and cosmetic problems is distributed similarly within conditions, also in case of the unique problems identified in each condition. However, we did find the types of identified problems to be similar between the coaching and traditional conditions while the active listening condition mainly led to problems concerning unclear information. In terms of problem severity, we found that the all think-aloud protocols revealed twice as many critical, serious and cosmetic problems as the silent condition.

## 5.2 Implications for Usability Practitioners

The most interesting result for usability practitioners is that the think-aloud protocols had only limited influence on user performance and satisfaction, but compared to the silent condition, they facilitated identification of the double number of usability problems with the same number of test subjects. Recruiting test subjects is a major task in usability evaluation, thus it is important to know how test subjects can be used most effectively. Furthermore, we found it difficult to interpret the causes of problems, which were observed through a lesser richness in problem descriptions.

Our results cannot be used to suggest practitioners to use a specific protocol. For example, the proponents of the Active listening protocol have argued that the Traditional protocol is unnatural. However, our results do not support this. The two protocols facilitate identification of a similar number of usability problems and richness in problem descriptions.

There were some interesting differences in the usability problems that were identified. The Traditional protocol revealed more usability problems in the frustration category, the Coaching protocol revealed more problems identified through test monitor intervention, and the Silent conditions mainly found usability problems identified when users made errors.

The results demonstrate that no single protocol version is superior overall; each has strengths and weaknesses. Yet in the silent condition, we could not capture verbalised problems, because the test subjects were silent, so the comparison is limited by this.

## 6 Conclusion

We have conducted an empirical study comparing three think-aloud protocols and a silent control condition. The study assessed the usability problems identified in each condition based on a systematic video analysis. We found that the differences between the three think-aloud protocols were limited. Contrary to the general emphasis on the Coaching protocol, we have no indication that it is superior for identifying usability problems. In fact, there were some aspects where the Traditional protocol performed surprisingly well. Overall, the think-aloud protocols are clearly superior to the silent condition; with the same number of test subjects, they revealed the double number of usability problems compared to the silent condition, and this applied across all levels of severity. Part of the study replicated a previous study on usability measures [19]. Here, we found only a few differences between the protocols. There was an indication that efficiency was higher in the Silent condition and that satisfaction with the website was higher for Coaching and Traditional. These limited effects do not contradict the literature as previous studies point in various directions.

Our study has a number of limitations. First, identification of usability problems is not as mechanical as determining usability measuring. Second, the specific experimental design impose limitations on our results, especially with our role as analysts, the moderator/analyst separation, and the physical moderator/user separation, but these were necessary to reduce potential sources of bias.

We deal with usability evaluation and the process of identifying usability problems. This is a controversial topic in HCI where some question the whole approach and its relevance to practice. Contrastingly, others maintain that within certain limits such studies produce relevant results. In order to achieve that, it is not possible to directly copy approaches from practice in the laboratory.

Our study points to interesting directions for future work. Most importantly, the consequences for and approaches used in practice should be studied extensively. It is also highly relevant to replicate the study with other types of systems.

**Acknowledgments.** We are grateful to Lise Tordrup Heeager, Rasmus Hummersgaard, Rune Thaarup Høegh Mikael B. Skov, Henrik Sørensen and the 43 users who helped us in the study.

## 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. In: Proceedings of Conference on Human Factors in Computing Systems 2007 (CHI 2007), pp. 1405–1414. ACM Press, New York (2007)
2. Boren, T., Ramey, J.: Thinking aloud: Reconciling theory and practice. *IEEE Trans. Prof. Commun.* **43**(3), 261–278 (2000)
3. Bruun, A., Gull, P., Hofmeister, L., Stage, J.: Let your users do the testing: a comparison of three remote asynchronous usability testing methods. In: Proceedings of Conference on Human Factors in Computing Systems 2009 (CHI 2009), pp. 1619–1628. ACM Press, New York (2009)
4. Cohen, J.: Statistical power analysis for the behavioral sciences, 2nd edn. Lawrence Erlbaum Associates, Hillsdale (1988)
5. Dumas, J., Redish, J.: A Practical Guide to Usability Testing. Intellect Press, Portland (1999)
6. Ericsson, K.A., Simon, H.A.: Protocol Analysis: Verbal Reports as Data, revised edn. MIT Press, Cambridge (1996)
7. Gray, W.D., Salzman, M.C.: Damaged Merchandise? A review of experiments that compare usability evaluation methods. *Hum. Comput. Interact.* **13**(3), 203–261 (1998)
8. van den Haak, M.J., de Jong, M.D.T., Schellens, P.J.: Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue. *Behav. Inf. Technol.* **22**(5), 339–351 (2003)
9. Henderson, R.D., Smith, M.C., Podd, J., Varela-Alvarez, H.: A comparison of the four prominent user-based methods for evaluating the usability of computer software. *Ergonomics* **38**(10), 2030–2044 (1995)
10. Hertzum, M., Hansen, K., Anderson, H.: Scrutinising usability evaluation: does thinking aloud affect behaviour and mental workload? *Behav. Inf. Technol.* **28**(2), 165–181 (2009)
11. Hertzum, M., Holmegaard, K.D.: Thinking aloud in the presence of interruptions and time constraints. *Int. J. Hum.-Comput. Interact.* **29**(5), 351–364 (2013)
12. Hertzum, M., Jacobsen, N.E.: The evaluator effect: a chilling fact about usability evaluation methods. *Int. J. Hum. Comput. Interact.* **15**, 183–204 (2003). Taylor & Francis
13. ISO 9241-11 (1998) Ergonomic requirements for office work with visual display terminals (VDTs). Part 11: Guidance on usability. ISO (1998)
14. Krahmer, E., Ummelen, N.: Thinking about thinking aloud: A comparison of two verbal protocols for usability testing. *IEEE Trans. Prof. Commun.* **47**(2), 105–117 (2004)

15. McDonald, S., Edwards, H., Zhao, T.: Exploring think-alouds in usability testing: an international survey. *IEEE Trans. Prof. Commun.* **55**(1), 2–19 (2012)
16. Molich, R.: User-Friendly Web Design (in Danish). Ingeniøren Books, Copenhagen (2000)
17. Nielsen, J.: Usability Engineering. Academic Press, Cambridge (1993)
18. Nørgaard, M., Hornbæk, K.: What do usability evaluators do in Practice? An explorative study of think-aloud testing. In: Proceedings of DIS 2006, pp. 209–219. ACM Press, New York
19. Olmsted-Hawala, E.L., Murphy, E.D., Hawala, S., Ashenfelter, K.T.: Think-aloud protocols: a comparison of three think-aloud protocols for use in testing data-dissemination web sites for usability. In: Proceedings of CHI 2010, pp. 2381–2390. ACM, New York
20. Rhenius, D., Deffner, G.: Evaluation of concurrent thinking aloud using eye-tracking data. In: Proceedings of Human Factors Society 34th Annual Meeting, pp. 1265–1269 (1990)
21. Rubin, J., Chisnell, D.: Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests. Wiley, Hoboken (2008)
22. Skov, M.B., Stage, J.: A conceptual tool for usability problem identification in website development. *Int. J. Inf. Technol. Web. Eng.* **4**(4), 22–35 (2009)
23. Wixon, D.: Evaluating usability methods: why the current literature fails the practitioner. *Interactions* **10**(4), 29–34 (2003)
24. Woolrych, A., Hornbæk, K., Frøkjær, E., Cockton, G.: Ingredients rather than recipes: A proposal for research that does not treat usability evaluation methods as indivisible wholes. *Int. J. Hum.-Comput. Interact.* **27**(10), 940–970 (2011)
25. Wright, R., Converse, S.: Method bias and concurrent verbal protocol in software usability testing. In: Proceedings of Human Factors Society 36th Annual Meeting, pp. 1220–1224 (1992)
26. Zhao, T., McDonald, S., Edwards, H.M.: The impact of two different think-aloud instructions in a usability test: a case of just following orders? *Behav. Inf. Technol.* (2012)

# An Observational Study of How Experienced Programmers Annotate Program Code

Craig J. Sutherland<sup>(✉)</sup>, Andrew Luxton-Reilly, and Beryl Plimmer

University of Auckland, Auckland, New Zealand

{c.j.sutherland, a.luxton-reilly,  
b.plimmer}@auckland.ac.nz

**Abstract.** This study investigates how and why experienced programmers annotate program code. Research has shown that marking up prose with a pen is an invaluable aid to reading for understanding. However program code is very different from prose: there are no studies on how programmers annotate code while reading. We asked experienced programmers to read code printed on paper and observed their annotation practices. We found the main reasons for annotating code are to assist with navigation and to record information for later use. Furthermore, we found annotation practices that are hard to replicate in current standard Integrated Development Environments. This suggests that support for digital ink annotations in programming tools may be useful for comprehending program code.

**Keywords:** Freeform annotation · Reading code · Understanding code · Observational study

## 1 Introduction

Reading code for understanding is an important part of maintaining computer applications [1]. Programmers can spend a significant portion of their time trying to understand code [2]. But reading for comprehension is not limited to software development, people read for comprehension for many different tasks [3]. What is unique about reading program code is it is almost exclusively read on a computer screen.

Programmers will often use tools when trying to understand code [4]. These tools can assist by providing alternate ways of navigating and different visualizations of the code. They attempt to reduce the amount of work the programmer has to do so the programmer can focus on comprehension [4]. However research has found that programmers do not always use these tools. Instead they prefer to rely on standards, experience and communications with other programmers [4].

Research into reading comprehension in other domains has found ink annotations are beneficial for improving understanding and retention of what has been read [5–9]. Annotating with a pen is a subconscious task: the reader annotates without interrupting the reading process [10]. Freehand ink annotation is more effective than the types of text annotation available in word processors [10, 11].

Code today is written and read directly on computers with programmers very rarely printing out code. As ink annotations provide benefits in other domains we speculate

that programmers may also get similar benefits. But there are no previous studies on how programmers annotate code. We hypothesize that when programmers read code they would add ink annotations if it were possible.

In this study we investigated how experienced programmers annotate when reading for comprehension. We analyze the nature and the purpose of their annotations. From this we compare their annotations to standard programming environment support and suggest ways in which ink annotations in Integrated Development Environments (IDEs) may be beneficial.

## 2 Background

### 2.1 Annotations in Reading

Annotations allow the reader to form a conversation with the text as they read. This changes the reader from being a passive receiver to being actively engaged in the text. Being actively engaged in the text helps people to understand and remember information [10]. Previous studies have found that annotation is intrinsically linked with reading [12, 13].

Annotations provide value over notes on separate paper by being close to the associated context. Having the context is especially useful when the reader is working through problems as it allows the reader to visualize the annotations and context simultaneously [14]. There is also evidence that the physical location of the annotations assists with finding them again [15]. People know roughly where they added something and will quickly flick to that location. In addition hand-written annotations stand out from the underlying context which also makes it easier to find them again [15].

Another strategy where annotations assist is with learning information by summarizing it [7]. The reader summarizes key points as they occur. Then during re-reading these summaries reduce the amount of re-reading needed. Annotations provide value over separate notes because they are situated in context. If the reader needs more information they can directly refer to the original text [16].

Different styles of reading have different characteristics. Some styles are characterized by a requirement of non-linear reading [17]. Non-linear reading is more demanding cognitively than linear reading. Readers need to hold what they are reading in memory at the same time as finding the next relevant section. In this context annotations reduce the mental workload [17].

### 2.2 Nature of Programs

Program code is very different from other forms of text. Program code is non-linear: programmers do not read sequentially from start to finish but trace paths through the code instead [18]. Code is highly structured: it follows rules that allow computers to execute it, and conventions that help programmers read it. Large programs are often split across multiple files. Tracing the flow of execution can involve swapping between multiple files and/or multiple locations within each file.

These differences make reading code very different from other forms of reading. Tools can compensate for these differences by assisting programmers with comprehending code. Examples of support provided by tools include alternate visualizations (e.g. graphs and diagrams), debugging support and syntax highlighting [19, 20]. There are also tools that allow people to draw diagrams and annotate their code [21, 22].

Navigating through code is an important task for programmers. If they cannot find a section then they may not understand the surrounding code. Navigation tools are common in IDEs. Common tools include search, project explorers and quick links. These tools can help programmers quickly find their way around code.

There are very few tools that allow freeform digital ink annotations and none of these are commercially available [18, 21, 23, 24]. This limitation means there have been no studies on how programmers might annotate code if there were no imposed constraints. There is some anecdotal evidence that annotations may help with navigation and comprehension [23]. There is also anecdotal evidence that the number of annotations decreases as the size of the program increases [25].

### 3 Methodology

For this study we observed experienced programmers reading a program they had not seen before. We investigated the following research questions:

- How do experienced programmers annotate code?
- Why do they annotate code?

#### 3.1 Participants

There were thirteen participants in this study. All were programmers with at least five years programming experience and they all had prior experience with C#. They all currently worked in a commercial environment or had previously worked in one. All but one participant was male.

Experienced programmers were chosen for this study to reduce any confounding effects due to prior experience. It is assumed they know how to read code already.

#### 3.2 Task

The participants were all asked to read a short block of program code. Prior to the reading task they were told that they would have to explain how the code works to a less experienced programmer. The rationale given for the task was that a new graduate would be taking over responsibility for the code and needed assistance with understanding the code. They were not told to annotate the code. If asked they were told they could do anything with the paper; including drawing or writing on it.

The code consisted of six files from a larger system written entirely in C#. The code in this study was responsible for initializing communications between modules in the

system. All relevant code was provided. The code was printed for the participants to read. The printed code used 14 pages of paper.

Paper was chosen as a medium to reduce the effects any particular tool might have. There are some tools that allow freeform annotations on screen but these are currently not robust and programmers are not used to using them [18, 21, 23, 24]. In contrast reading on paper is a common task for most people.

Each participant was given a copy of the code, some extra blank pieces of paper, a highlighter and four different colored pens (red, blue, black and green).

The participant was seated at one side of a table. At the other side of the table a video camera was setup to record what the participant was doing. The camera was focused on the participant's hands and the pieces of paper. The investigator sat behind the camera and ensured it was correctly focused and adjusted as needed during the task. The investigator also took notes during the task.

### 3.3 Procedure

Each participant performed the task in a controlled environment. Prior to the reading task each participant was welcomed to the lab, had the process explained to them (including gaining consent) and filled in a short questionnaire on their programming and reading background.

Each participant was given 45 min to read the program code. They could finish earlier if desired. While reading each participant was video recorded and observed. We answered questions about the process (e.g. can I write on this paper?) but not about the code (e.g. what does this class do?) At the end of the reading time their paper was collected and the purpose of the study was explained.

After the reading task there was a short interview about their annotations. This involved looking at the annotated paper and asking the participant to explain why they had made each annotation. This included the importance of the annotation and how it fitted into their approach to understanding the code.

### 3.4 Analysis

After each participant finished the data was coded. We went through each page and counted the number of annotations and recorded the following attributes about each:

- Location
- Type of annotation
- Reason for adding annotation.

Location was a choice from within code or left | right | top | bottom of code. This was based on where the majority of the annotation was.

The type of annotation was based initially on the annotations described by Marshall [26]. When a new annotation type was found the list was expanded to include it. After coding the list was reviewed and similar types of annotations were combined.

During the interview the investigator asked the participant the general purpose of each annotation. This was then summed up in a few words (e.g. "question", "highlight

for later”, “possible bug”, etc.) The reasons were then consolidated into a list of basic reasons based on the participant’s intention when the annotation was added.

These attributes were then summarized and collated with the answers from the questionnaires.

## 4 Results

The majority (9 out of 13) of the participants currently used C#. All of these participants reported using it daily. The other participants all had some experience with C# and were confident that they could read a short program in it.

Most participants reported reading code on a daily basis. Reasons for reading included debugging, learning, reviewing other people’s code and general development (e.g. extending an existing codebase). Most participants read their own code on a daily basis. The two participants who did not read their own code on a daily basis did read other people’s code at least weekly. While most participants reported reading code on screen daily only one participant reported regularly printing off code for reading. All other participants said they rarely printed code to read on paper.

During the study 12 of the participants added annotations or notes. Only one participant (P05) did not add any annotations or notes. In the post-observation interview we asked about this. He reported he never makes any annotations and very rarely takes notes when using pen and paper. His reason for this was it was the way he was trained. He was excluded from the remaining analysis as an outlier.

### 4.1 How Programmers Annotate

In total the twelve participants added 267 annotations (see Table 1). Ten participants added annotations on the code; the other two participants only wrote on separate pieces of paper. The average number of annotations added by each participant was 33. The number of files annotated ranged from one to all six. On average each participant added eight annotations per file.

The following types of annotations were identified:

- Underline: a line drawn underneath text
- Scratch-out: a line drawn through text
- Highlight: a line drawn through text with the highlighter
- Enclosure: a circled block of code
- Margin bar: a vertical line, typically drawn in the margins
- Brace: a } like annotation spanning multiple lines of code
- Connector: a joining line between two elements (e.g. a block of code or another annotation) without an arrow
- Arrow: as above but with an arrow head
- Text: one or more characters (e.g. letters, numbers, punctuation marks)
- Drawing: a diagram or picture
- Dot: a small mark or dash.

**Table 1.** General Annotation Details

Participant	Number of annotations added	Number of files annotated	Average annotations per file
P01	44	6	7
P02	36	5	7
P03	21	2	11
P04	76	6	13
P06	31	4	8
P07	12	4	3
P08	0	—	n/a
P09	29	2	15
P10	0	—	n/a
P11	5	2	3
P12	9	1	9
P13	4	2	2

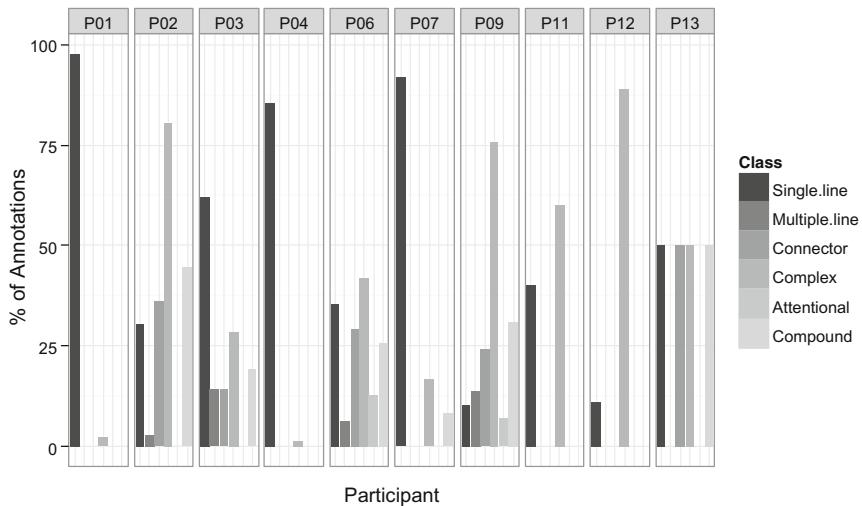
These annotations were classified into five categories:

- Single line: these span a single line and are associated with code or other text. Underlines, highlights and scratch-out all fall into this category.
- Multiple line: these span multiple lines of code but do not contain explicit meaning beyond their location. Enclosures, margin bars and braces fall into this category.
- Connector: these join two or more items together. One common use for a connection annotation is to associate a text annotation with a segment of code. Arrows and connectors fall into this category.
- Complex: these have explicit meaning associated with them, although maybe only to the original annotator. Text and drawings are both in this category.
- Attentional: these are indirect signs of the participant’s attention. They are temporary and not intended for future re-use. Dots fall into this category.

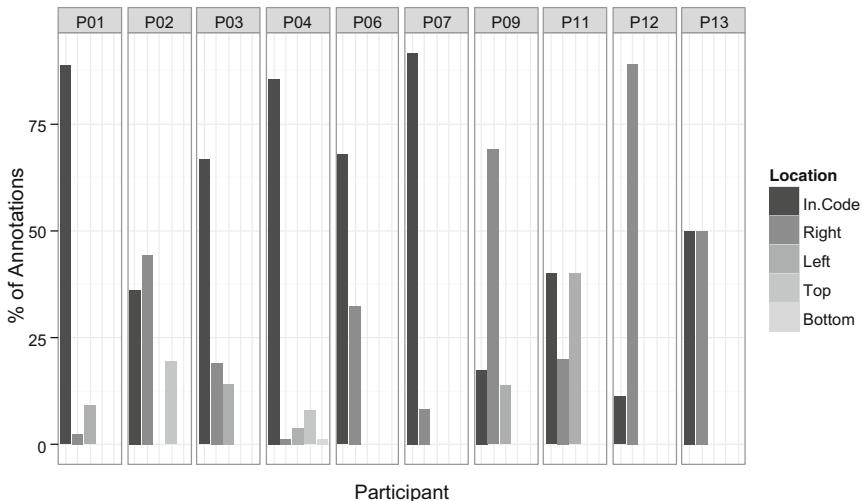
In addition some annotations consist of multiple annotation types: these are compound annotations. The participant added these together at the same time and considered them as a single annotation. A common example is a connector and text together.

The annotations added by each participant are shown in Fig. 1. While there are no common patterns among the participants in this study there are two general trends. Four of the participants (P01, P03, P04 and P07) preferred mainly single-line annotations. Four of the remaining participants (P02, P09, P11 and P12) preferred mainly complex annotations. Of these annotations they are predominately text notes.

The majority of the annotations were added within the code (65 %). The next most common location is to the right of the code (24 %) and then to the left or top of the code (6 % and 5 %). Only one participant added an annotation below the code – this involved copying some lines of code from the next page. Figure 2 shows where each participant added the annotations.



**Fig. 1.** Percentages of annotation types added on the source by classification per participant. Percentages are percentage of total annotations added by the participant. Compound annotations are a separate classification – therefore the totals may add to more than 100 %.



**Fig. 2.** Percentage of annotations added to each location per participant. Percentages are percentage of total annotations added by the participant.

There is a relationship between the classification of annotation and its location. Single line and attentional annotations mainly occurred within the code. Multiple line and connector annotations mainly occurred to the left or the right of the code. Complex annotations occurred at any location.

## 4.2 Why Programmers Annotate

Initial data entry identified 39 different reasons for adding the annotations. During the analysis we identified that many reasons were very similar in nature. Therefore similar reasons were grouped together resulting in eleven basic reasons which fall into four categories (see Table 2).

We grouped these basic reasons into four categories based on the main intent. The first category of reasons, navigation, is for helping the participant find their way through the code. These typically speed up the process of re-finding information. The second category, working information, is where the participant is recording things they have found about the code. This is information they think will be useful later on. The third reason, information for sharing, is similar but intended for someone else. The final reason, other, is for the more uncommon annotation reasons.

**Navigation Annotations.** The most common reason for adding an annotation is to emphasize something for future use. This marks an item that the participant is thinking of returning to later – even if they never do. Common examples of this are: underlining a method call or name, circling a section of interesting code and highlighting where variables are defined. These annotations were generally single-line annotations.

There were two participants who extensively annotated for this reason (P01 and P04). They both said they were trying to build a model of the code. P01 said he was trying to replicate the solution explorer in Visual Studio (the solution explorer lists all the classes and methods in a set of code files). He went quickly through the code initially to highlight most of the method names. Then he went through the code a second time using the highlighted method names to trace what was happening. In contrast P04 used the highlighted method names as a form of backtracking. When he started in a method he would highlight the method name. Later he could easily return to this method by scanning through the paper for the highlight.

Other participants who emphasized for future reference were more selective. They only emphasized the elements of interest. There were two main motives: emphasizing hard-to-find elements and commonly referenced elements. Both motives were to help the participant re-find them later.

Another reason is to add a reference. These annotations link two or more sections of code together. The participants added these when they thought they would need to move quickly between two segments of the code. One example of this reason is matching letters to the side of the code. The participant was able to quickly scan through the code and find the matching reference as needed. These annotations were mainly text like a symbol or the name of the method. During the interview participants said the references were not always used but when used were valuable especially for hard-to-find segments of code.

By emphasizing code structure the participant was trying to make it easier to move between different files. In all cases the participants added these at the top of the first page in the file. These annotations were either text or a highlight. The highlight or text was always positioned so the participant could see it quickly.

Emphasizing a significant feature had two categories: important code and hard-to-find code. These differ from other navigation reasons in one significant aspect: the

**Table 2.** Consolidated list of reasons why annotations were added.

Reason	Description	Number	Percentage <sup>a</sup>
<b>Navigation</b>			
Emphasize for future	The annotation marks a feature of the code to be reviewed later	117	44 %
Add reference	A reference to a different section in the code	21	8 %
Emphasize code structure	Highlight a structural element in the code	17	6 %
Emphasize significant feature	A section of the code that needs further investigation	9	3 %
<b>Working Information</b>			
Record working notes	Inline notes on what is happening in the program	36	13 %
Record question	A query about something in the code	36	13 %
Correct previous annotation	A correction or update to a previous annotation	6	2 %
<b>Information Sharing</b>			
Record a needed change	An area of the code that needs to be changed	11	4 %
Emphasize example	An example of another section of code is called	2	1 %
<b>Other</b>			
Unknown	The participant was unsure as to why they added the annotation	6	2 %
Attentional	An incidental mark as a result of the user's attention	6	2 %
<b>Total Number of Annotations:</b>		<b>267</b>	

<sup>a</sup>Due to rounding these percentages do not add to 100 %.

participant expects they will implicitly remember the rough location of the annotation. The annotation speeds up finding the location again. One participant mentioned this is like adding a post-it note to the page.

**Working Information Annotations.** We identified three reasons for annotations that were for recording information. These reasons were: working notes; questions and corrections. All three reasons were an attempt to reduce the cognitive workload while reading.

Working notes are a description about what is happening in the code. The participant would add a working note when they figured out something and wanted to remember it later. The notes are a way of offloading this information onto paper to reduce their mental workload. These are typically text comments although they might only make sense to the writer. Occasionally a participant would return to their working notes and make a correction if they realized a comment was incorrect.

A question annotation is where the participant wants to find out some more information. Questions are different from working notes because the participant is seeking an answer. In contrast working notes are what the participant already understands. Unlike working notes question annotations have a wide variety of forms. They could be as simple as an underline or asterisk next to some code right up or a complete written question. During the interview we asked for more details about these question annotations including what the underlying question was and whether it was answered later.

We found three types of question: implementation, functional requirements and language features. A question on the implementation is why the code was implemented a certain way. For example, one participant was interested in why there were two methods with the same name but different parameters. A question on the functional requirements was when the participant saw something that was not clear if it matched the described functionality of the system. For example, one participant questioned whether a method had the correct logic. Finally a question on a language feature was when the participant saw something they did not understand about the language or base libraries. For example, one participant queried about how a lambda was coded.

An implementation question was typically answered by the participant in their reading when they understood the code better. But only one participant updated these annotations with the answer. Most of the participants said they just remembered the information in their head. The annotation reminded them of the question and answer. Questions about the functional requirements were sometimes answered when the participant read other parts of the code. None of the participant updated these annotations although some added a reference to the question. One participant described these as questions to take to the business owner about whether the code is correct. The final type were questions that would be answered by researching the feature: looking online, asking someone or looking up some form of reference materials. None of the participants updated these annotations during the session.

A correction annotation is where the participant has returned to a previous annotation and updated it. Corrections always involved scratching out part or all of an annotation. In two instances the participant also added additional alternate text.

**Information Sharing Annotations.** Participants also added annotations to record something that needed changing. Examples of these annotations included hard-coded values, unclear variable or method names and bad coding practices. Most were enclosures; there was only one text annotation in this category. The participants who added these annotations stated they only had to look at the code to remember why they added the annotation. This implies that these annotations are a form of offloading from memory onto the paper.

The least common reason was emphasizing an example. Both example annotations were where the participant thought the code was a very good example for a junior programmer to see.

**Other Reasons.** Attentional annotations are incidental marks of the participant's current focus. During the task the participants often pointed to the code with the pen currently in hand. Occasionally the participant would make contact with the page. During the interview the participants who made these annotations mentioned they were accidental and they did not care about these annotations.

Finally, we added a reason of unknown. These are when the participant did not remember why the annotation was added. However this was uncommon: they only failed to remember the reason six times out of 267 annotations.

### 4.3 Separate Notes on Paper

In addition to making annotations on the code nine of the participants made notes using separate pieces of paper. Three of these participants used two pieces of paper while the rest used a single piece. Only one of the participants added notes to both sides of the piece of paper. This resulted in a total of thirteen pages of notes. Of these thirteen pages four were text only. The remaining pages contained diagrams and text (see Fig. 3).

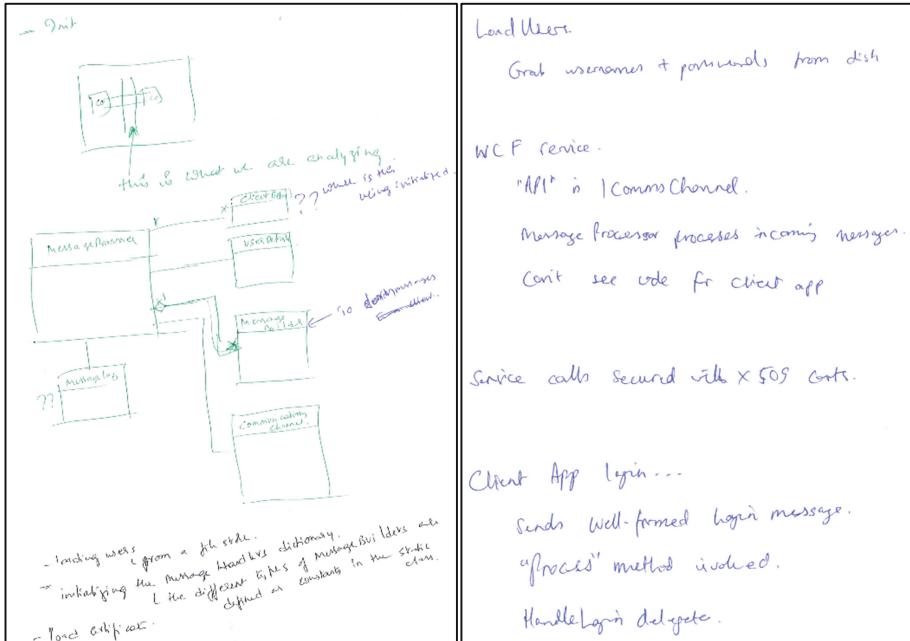
The most common usage for notes is class diagrams and flow diagrams. In class diagrams the participant is drawing a graphical representation of the classes. There were six class diagrams drawn. A flow diagram shows a flow of either information or control through the code. There were five flow diagrams drawn.

The class diagrams typically included the interfaces defined and relationships to other classes. These were mainly drawn at an overview level. Only one of the diagrams contained property and method members. In addition these members were a subset of the full members for the class. One of the diagrams was text only; the rest contained text and graphics elements. These ranged from simple arrows connecting classes and/or interfaces to boxes similar to UML format. Two of the diagrams contained textual notes in addition to the diagram (see Fig. 3). These were notes explaining the purpose of some of the classes.

The flow diagrams were more varied. Three diagrams worked at a high level (i.e. server and client) and showed some of the flows between these components. One of these had the relevant classes written for each component, another had the methods and the third was a conceptual diagram. The other flow diagrams were all based at the class level. They had some of the classes in the code and the messages that were passed between them. Two flow diagrams had textual notes. One set of notes listed the data structure being passed around; the other listed some background information to the data flows.

Two pages contained notes of findings (see Fig. 3). These findings were information the participant thought relevant to share with the junior programmer. Both of these pages were text only. They were both grouped lists. The first line in each group was the heading and the remaining lines were indented.

We matched these pages with two of the basic reasons: working information and information for sharing. During the interviews the participants stated the information was either for themselves or for sharing with the junior programmer. Three of the participants started writing the notes with the intention of sharing. The other six participants intended the notes to be only for themselves; these notes were working information. Three of these participants stated that they would share their notes with someone else. However they stated they would not directly share the working notes as they did not consider the information complete. They expected to either sit down with the person reading it or to revise it into a form that could be shared.



**Fig. 3.** Examples of note pages: left page (from P12) shows a class diagram with added notes; right page (from P09) shows textual notes for sharing with another programmer.

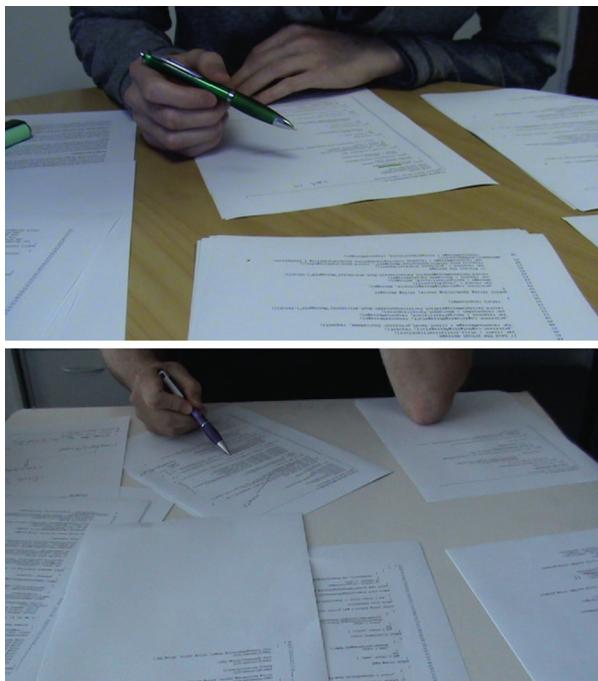
The time when the participants wrote the notes also varied widely. Two of the participants started writing notes very soon in the reading process. These participants did not make any annotations on the code; instead they used the separate pages for writing information. Five of the participants started writing notes later in the process. These participants used a combination of annotations on the code and separate note pages. The remaining two participants annotated the code as they read. Near the end of their reading they then went through and wrote notes on the separate paper.

The reason for the notes was linked to when the participant started writing them. The two participants who wrote the notes at the end of the session wrote the notes specifically for the junior programmer. In contrast the two who started writing at the start of the session wrote the notes for their own understanding. The remaining participants were split between writing for the other person and for themselves. These were also the participants who changed the intention of the notes.

#### 4.4 Other Observations

In addition to the observations about the annotations there were some other interesting patterns. These are not specific to annotating code but reading code on paper in general.

One of the first things eleven of the participants in the study did was split the code into the separate files (most files contained a single class, one contained two). They



**Fig. 4.** Examples of paper spread around the desk.

were originally given a single stack of paper containing all the code: they took this stack and leafed through to find how many different files there were. Then they separated the files into separate stacks of paper.

All the participants spread these stacks of paper around the desk (see Fig. 4). They reported a variety of reasons why they did this: to keep the files separate, to gain an overview of the code, to know where the various files are (for moving between them). Some of the participants combined this separating of pages with annotations. For example, P02 wrote the class names on the top of the first page in each class so he could quickly find them again.

Seven of the participants combined the separation of pages with spatial positioning. Once they had separated the piles they tended to keep them in the same positions. They would pick up a pile to read or search through it and then return it to the same position. For four of these participants this was a conscious behavior: they had deliberately made the decision to do this when separating the piles. One participant (P11) had even arranged the location of the piles based on their sizes. The underlying rationale was the bigger piles were more likely to be used. So he placed them in locations where he could quickly reach them. The other three participants used spatial positioning without having made a conscious decision on it. One participant, P05, stated that he did not originally intend to keep them in the same locations but midway through the read task found it was easier to find things if he returned them to the same location.

## 5 Discussion

In this study all but one participant added annotations to the code or wrote them on a separate piece of paper. This was expected as the environment was impoverished compared to what is available in most IDEs. There are a number of different reasons given for adding annotations.

The frequency of some annotation types are similar to what has been reported in other studies (e.g. [27]). Underlines and highlights had similar frequencies to what has been reported before. This may be because they are so simple and easy to add.

One annotation type that is more frequent in this study is compound annotations. The majority of the compound annotations consist of a connector plus another type of annotation. We posit this is because annotations need to be associated with specific lines of code. If the participant did not need a precise anchor they would not use a connector but would be simply a sidebar as is seen in prose annotation.

The majority of annotations were added to help the participant navigate the code. This is similar to what other research has shown for when programmers read code [1, 2, 4]. Our participants used a range of strategies to mark specific sections of code (such as highlights and margin bars) and also cross-reference marks to indicate connections between code on different pages. For example most participants highlighted some method names. This selective highlighting, while similar to IDE color syntax highlighting, is more specific in that the participants highlighted only some method names.

Another common form of navigation annotations was for backtracking. Again this is functionality that IDEs already provide. But IDE navigation support is generic where annotations are specific. The readers annotate to focus on what they are interested in and to reduce the workload by place marking. This type of annotation is seen in other studies [15, 17] but the importance of navigation in code comprehension makes it more critical for code understanding. Digital ink annotation in IDEs could be a valuable aid to support code navigation.

Computers can, and do, generate a lot of navigation information but they do not know what the reader is interested in. Therefore the reader may take false routes to find what they need, sometimes even backtracking to revisit previously scanned code [4]. While annotations cannot help find unvisited code they can help reduce the clutter of backwards navigation. A combination of computer generated navigation links and user ink annotation may be optimal.

The other main reason was for recording information. Offloading information to paper reduces the amount of cognitive workload the reader needs to perform [15]. P08 mentioned that one of his challenges was trying to remember everything. This reduction in cognitive load and abstraction that ink annotation affords is consistent with other studies [10, 11] suggesting that ink annotation inside code editors could offer similar benefits.

Annotations on the code and on the separate pieces of paper both help to reduce the cognitive load. However there appear to be different benefits between the two. Annotations on the code were shorter and more cryptic than on the separate notes. In contrast the diagrams on the separate paper would be more difficult to add to the code because of space constraints and also they tend to be an overview rather than location

specific. However the separate notes were artifacts that the participants mentioned they would want to keep. These results corroborate the ideas of Lichtschlag, Spychalski and Bochers [22] who suggested that hand-drawn sketches assist programmers in orientating in a codebase.

The spreading out of the paper on the table is an interesting observation that could benefit from further research, in particular is the importance of spatial positioning. There are tools that allow graphical representations for code (e.g. [19, 20]). These tools should maintain the spatial positioning of the elements to help programmers remember where things are. Large screen or multiple screen systems may also be beneficial by providing a larger space for programmers to work. The programmer could move the code files into positions that they find useful. The current file would be moved into a prominent location for work. Then when finished it would be returned to its previous location. If the programmer needs it later they can easily find where it is based on its location. We are not aware of any research into this type of programmer support.

This study is part of a larger project looking at utilizing freeform digital ink annotations in an IDE. In code editing tools the main way to record information is by adding comments in the code. Previous studies comparing textual annotations vs. freeform annotations found people prefer freeform annotations [10, 11]. O’Hara and Sellen suggested the most likely reason for this is freeform annotations do not interrupt what the reader is thinking about [10]. Therefore freeform annotations may be useful for storing information against the code without reducing the reader’s capacity for comprehending what they are reading. Second, annotations are not limited to text. Several of the participants drew diagrams of how they understood things. Studies have found that diagrammatic annotations provide value by keeping the diagram close to its context [14]. Finally, annotations stand out from the text. Freeform annotations look very different from the underlying document. Many of the participants in the study were able to quickly find things by just scanning through the code and looking for the annotations. This, when combined with spatial memory, allowed them to easily find something they had previously written.

## 6 Limitations

First, participants were given a specific task to do (read the program code for understanding). A different task may have resulted in different annotation patterns (e.g. marking an assessment, adding unit tests). Second, most of the participants used IDEs that provide a variety of tools to assist with understanding. Some of these tools might be replicable on paper (e.g. an index of all the method locations). If these tools were available then the participants may have used different types of annotations. However we contend that the selective nature of annotations may, in some circumstances, be better than generic functionality. Third, the programming language may have an effect on the types of annotation. Some languages are more difficult to read: a programmer reading program code in these languages may add different types of annotations. Finally, this study only had 13 participants. Including more participants might make some of the patterns observed more obvious. However this study does provide some insight on how annotations could be useful in an IDE.

## 7 Conclusions and Future Work

This study investigated how and why experienced programmers annotated code on paper. Previous work has shown it is possible to combine freeform ink annotations within a code editor [21, 24] but there was no evidence of if, how and why programmers would annotate code with digital ink.

The results of this study indicate that somewhere to record information would assist programmers comprehend what they read. The ability to annotate on code with free-form ink may be useful if the functionality were available. Separate blank pages may also be beneficial; especially if there is some way to link the notes to the code (see [22]).

This study shows that programmers use annotations to assist with navigation, record information as working notes and to remember information for sharing. While some of this functionality is currently available in code editing tools, user specified particular navigation paths, which were frequently used in our study, are not supported in IDEs. Furthermore, freeform ink annotations provide an alternate avenue that reduces the cognitive work needed when reading code. This is because ink annotations are quicker and easier to add compared to text-based annotations, allow greater expressiveness and stand out from the code because they are visually distinct from the text.

It would be interesting to compare how participants used an IDE for the same task. Comparing the two toolsets would allow a comparison of what functionality is currently available and what could be provided in future.

Future lines of investigation include:

- How would programmers annotate code if an IDE supported freeform annotation?
- How can freeform annotations assist navigation?
- How could spatial layout of files assist comprehension?

**Acknowledgment.** We would like to thank all the participants in our study for time they gave us.

## References

1. Sillito, J., De Volder, K., Fisher, B., Murphy, G.: Managing software change tasks: an exploratory study. In: International Symposium on Empirical Software Engineering 2005, pp. 23–32 (2005)
2. Singer, J., Lethbridge, T., Vinson, N., Anquetil, N.: An examination of software engineering work practices. In: CASCON First Decade High Impact Papers, pp. 174–188. IBM Corporation, Toronto (2010)
3. Sellen, A.J., Harper, R.H.R.: The Myth of the Paperless Office. MIT Press, Cambridge (2002)
4. Maalej, W., Tiarks, R., Roehm, T., Koschke, R.: On the comprehension of program comprehension. ACM Trans. Softw. Eng. Methodol. **23**, 1–37 (2014)

5. Fowler, R.L., Barker, A.S.: Effectiveness of highlighting for retention of text material. *J. Appl. Psychol.* **59**, 358 (1974)
6. Hynd, C.R., Simpson, M.L., Chase, N.D.: Studying narrative text: The effects of annotating vs. journal writing on test performance. *Reading Res. Instruction* **29**, 44–54 (1989)
7. Simpson, M.L., Nist, S.L.: Textbook annotation: an effective and efficient study strategy for college students. *J. Reading* **34**, 122–129 (1990)
8. Wolfe, J.L., Neuwirth, C.M.: From the margins to the center: the future of annotation. *J. Bus. Technical Commun.* **15**, 333–371 (2001)
9. Ball, E., Franks, H., Jenkins, J., McGrath, M., Leigh, J.: Annotation is a valuable tool to enhance learning and assessment in student essays. *Nurse Educ. Today* **29**, 284–291 (2009)
10. O'Hara, K., Sellen, A.: A comparison of reading paper and on-line documents. In: *Proceedings of CHI 1997*, pp. 335–342. ACM, New York (1997)
11. Morris, M.R., Brush, A.B., Meyers, B.R.: Reading revisited: evaluating the usability of digital display surfaces for active reading tasks. In: *Proceedings of TABLETOP 2007*, pp. 79–86. IEEE (2007)
12. Adler, A., Gujar, A., Harrison, B.L., O'Hara, K., Sellen, A.: A diary study of work-related reading: design implications for digital reading devices. In: *Proceedings of CHI 1998*, pp. 241–248. ACM, New York (1998)
13. Hong, M., Piper, A.M., Weibel, N., Olberding, S., Hollan, J.: Microanalysis of active reading behavior to inform design of interactive desktop workspaces. In: *TABLETOP 2012*, pp. 215–224. ACM, New York (2012)
14. Jackel, B.: Item differential in computer based and paper based versions of a high stakes tertiary entrance test: diagrams and the problem of annotation. In: Dwyer, T., Purchase, H., Delaney, A. (eds.) *Diagrams 2014*. LNCS, vol. 8578, pp. 71–77. Springer, Heidelberg (2014)
15. Johnson, M., Nadas, R.: Marginalised behaviour: digital annotations, spatial encoding and the implications for reading comprehension. *Learn. Media Technol.* **34**, 323–336 (2009)
16. Glover, I., Xu, Z., Hardaker, G.: Online annotation – Research and practices. *Comput. Educ.* **49**, 1308–1320 (2007)
17. Crisp, V., Johnson, M.: The use of annotations in examination marking: opening a window into markers' minds. *Br. Educ. Res. J.* **33**, 943–961 (2007)
18. Priest, R., Plimmer, B.: RCA: experiences with an IDE annotation tool. In: *Proceedings of CHINZ 2006*, pp. 53–60. ACM, New York (2006)
19. Bragdon, A., Reiss, S.P., Zeleznik, R., Karumuri, S., Cheung, W., Kaplan, J., Coleman, C., Adeputra, F., LaViola Jr., J.J.: Code bubbles: rethinking the user interface paradigm of integrated development environments. In: *Proceedings of ACM/IEEE International Conference on Software Engineering*, pp. 455–464. ACM, New York (2010)
20. DeLine, R., Bragdon, A., Rowan, K., Jacobsen, J., Reiss, S.P.: Debugger canvas: industrial experience with the code bubbles paradigm. In: *Proceedings of the 2012 International Conference on Software Engineering*, pp. 1064–1073. IEEE Press (2012)
21. Sutherland, C.J., Plimmer, B.: VsInk: integrating digital ink with program code in visual studio. In: *Proceedings of AUIC 2013*, pp. 13–22. Australian Computer Society, Incorporation (2013)
22. Lichtschlag, L., Spychaliski, L., Bochers, J.: CodeGraffiti: using hand-drawn sketches connected to code bases in navigation tasks. In: *Proceedings of VL/HCC 2014*, pp. 65–68. IEEE (2014)
23. Chen, X., Plimmer, B.: CodeAnnotator: digital ink annotation within Eclipse. In: *Proceedings of AusCHI 2007*, pp. 211–214. ACM, New York (2007)

24. Lichtschlag, L., Borchers, J.: CodeGraffiti: communication by sketching for pair programmers. In: Adjunct Proceedings of ACM Symposium on User Interface Software and Technology, pp. 439–440. ACM, New York (2010)
25. Plimmer, B.: A comparative evaluation of annotation software for grading programming assignments. In: Proceedings of AUIC 2010, pp. 14–22. Australian Computer Society, Incorporation (2010)
26. Marshall, C.C.: Annotation: from paper books to the digital library. In: Proceedings of ACM International Conference on Digital Libraries, pp. 131–140. ACM, New York (1997)
27. Marshall, C.C., Brush, A.J.B.: Exploring the relationship between personal and public annotations. In: Proceedings of Digital Libraries, 2004, pp. 349–357. ACM Press, New York (2004)

# Around-Device Interactions: A Usability Study of Frame Markers in Acquisition Tasks

Fernando Garcia-Sanjuan<sup>1(✉)</sup>, Alejandro Catala<sup>1</sup>, Geraldine Fitzpatrick<sup>2</sup>,  
and Javier Jaen<sup>1</sup>

<sup>1</sup> ISSI/DSIC, Universitat Politècnica de València, Valencia, Spain

{fegarcia, acatala, fjaen}@dsic.upv.es

<sup>2</sup> HCI Group, Vienna University of Technology (TU Wien), Vienna, Austria  
geraldine.fitzpatrick@tuwien.ac.at

**Abstract.** Digital tabletops present numerous benefits in face-to-face collaboration environments. However, their integration in real settings is complicated by cost and fixed location. In this respect, building table-like environments using several handheld devices such as tablets or smartphones provides a promising alternative but is limited to touch interaction only. We propose instead another kind of “around-device” interaction (ADI) technique using the built-in front camera of these devices and fiducial frame markers, which presents advantages including better awareness and less interference. This paper contributes a first step in exploring the potential of this interaction technique by conducting a usability test comparing several ergonomic factors that may have an effect on the very first operation of the interaction: the acquisition of the marker.

**Keywords:** Around-Device Interaction (ADI) · Tablets · Fiducial markers · Frame markers · Multi-Display Environments (MDE) · Usability study

## 1 Introduction

Digital tabletops have been shown to be very suitable tools for use in collaborative environments [4, 6]. Their form factor improves workspace awareness [6], and their multi-touch capabilities allow simultaneous interaction. Together, this increases parallelism, allows more democratized access, and leads to an increased collaborative performance, which produces better results [6, 11]. Nevertheless, it is rare to see tabletops embedded in real settings. This is due to a number of disadvantages: their high cost; their limited workspace dimensions, which can only accommodate a certain number of participants; and the fact that their form factor complicates mobility in a way that, nowadays, if a digital tabletop is available, it is fixed to a single location and users are forced to move to a specific place if they want to engage in a collaborative activity around it. Ideally, it would be desirable for users to be able to form groups in an improvised way, in virtually any place, and of any size, using the devices they have with them to dynamically create a tabletop-like collaborative space. This allows us to take advantage of the benefits of tabletops in terms of awareness and parallel interaction, but using a different approach to tabletop working based on handheld devices. Devices such

as smartphones or tablets are becoming increasingly popular and will be in common use in the near future. The portability of these devices implies that it could be possible to build multi-display environments (MDE) to support co-located collaborative activities on a table-like setting by coordinating interaction across several devices put together on the same table. However, this scenario raises a new challenge. As Gutwin et al. [4] found out, in highly integrated collaborative scenarios, users often have the need to acquire elements that are out of reach. Therefore, since interaction with handhelds is usually carried out by touch contact, in a situation where several users are gathered around the same table, interference problems may arise related to simultaneous touch actions from different users on the same surface. The immediate solution comes by asking others to hand over the elements. However, this could interfere with the task another user is currently carrying out. Another possible solution is Around-Device Interactions (ADI) [8], where a user could interact with an out-of-reach tablet which is at the same time being manipulated via touch by another user, and without interfering with the latter's actions. We explore the potential of fiducial frame markers as an enabler of ADI.

Our end future goal is to design MDE collaborative environments around a table like the one shown in Fig. 1, where users bring their own devices and where they may use cards or tangible elements with attached fiducial markers to share objects (e.g., documents, game elements) or trigger reactive behaviors in a target surface on the table. The built-in camera in each device is used for the marker detection; therefore no additional hardware infrastructure is required. In this kind of setting, several interactive surfaces can be placed on a table at different distances, rotations, etc. Thus, the entry point at which users approach each surface can dramatically change, and so can the conditions surrounding the interaction itself. As such, before designing any complex ADI for these settings, the fundamental issue that must be addressed is the acquisition of the marker [13] (i.e., the initial step at which the fiducial marker is placed in the camera's field of view for it to be detected), and evaluating how usable this is under different conditions. The main aim of this paper then is to present an ADI technique that uses fiducial markers and to obtain the ground knowledge that will enable us to design a table-based MDE that uses it effectively. Specifically, we evaluate the usability of the acquisition phase of the proposed ADI through an in-lab study, focusing on the ergonomic conditions that are perceived as facilitators of this interaction.



**Fig. 1.** Hypothetical collaborative scenario with ADI using fiducial markers

## 2 Related Work

There are a number of different ADI techniques with varying degrees of flexibility and hardware complexity. For example, Hasan et al. [5] present AD-Binning, an ADI used to extend the workspace given by the small screen of a smartphone by including their surroundings. Users then interact with the smartphone by using a ring-shaped band on their fingers, and store/retrieve items to/from the extended space. Both the device and the ring are tracked by a complex hardware setup formed by eight external cameras. While this may be precise, it is not available to the common user and requires previous assembly and calibration. Avrahami et al. [1] also explore ADI by using cameras mounted on both sides of the tablets to track distinct objects and capture interactions in the space near the device. While this approach allows mobility and the possibility of forming groups of users virtually anywhere in an improvised way, it still requires a careful installation of external hardware. A simpler approach is by Kratz et al. [8], who attach external IR sensors on a smartphone screen, and allow ADI using hands. Their main purpose is to reduce the occlusion produced by touch contacts, which is a form of interference. Other works reduce the hardware complexity even further by making use of the integrated sensors in the tablets. Katabdar et al. [7], for instance, exploit the magnetic (compass) sensor of the device, and interact around it using magnets. Unlike optical approaches, this solution is more robust to occlusion. They also can shape the magnets in different forms, though the system is not capable of differentiating between different magnets since they do not have an encoded ID. This prevents its use in applications with multiple collocated users.

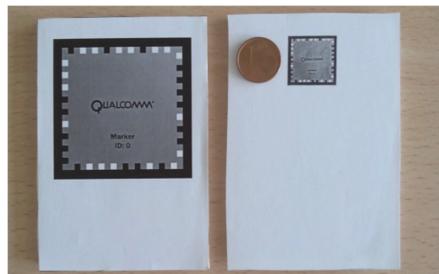
Although the previous examples enable ADI by augmenting tablets with external sensors, they are designed to be used by a single user and on one device. Probably for this reason most of them [1, 7, 8] require the interactions to take place close to the device. In a table-based MDE, this restriction could require that users lean over the table or even move towards the target device, which could be cumbersome and cause interference with interaction on the other devices.

Finally, Nacenta et al. [10] compare several interaction techniques, including ADI, in tabletop groupware in terms of how they enhance coordination. They include aspects like lack of interference, ease of transference of elements, and easy access to out-of-reach zones as important factors. Our work complements this previous work by considering other forms of interaction based on fiducial markers and studies the ergonomic, visual feedback, and marker size factors.

## 3 User Evaluation

In its simplest form, the proposed interaction consists of taking a card with the marker and bringing it close to a selected tablet for recognition. To study the usability of this, we identified seven ergonomic factors that may have an effect on this initial interaction phase: the *user posture* (sitting vs. standing), the *active hand* (non-dominant vs. dominant), the *marker size* (small vs. big, as shown in Fig. 2), the

*tablet position* with respect to the users (whether it is at their non-dominant vs. dominant side), the *tablet distance* to the users (near vs. far, depending on whether it is within arm's reach or not), the presence of a *visual feedback* on the tablet that allows the users to see what the camera sees (missing vs. present), and the users' *gender* (male vs. female). In presenting the results (see Fig. 3), the first condition listed is represented as "level -" and the second as "level +", e.g., sitting is the "-" level and standing is the "+" level for the *user posture* factor.



**Fig. 2.** Cards with the fiducial frame markers used in the experiment

### 3.1 Apparatus

The experiment was conducted with a  $120 \times 80$  cm Table 75 cm high in an environment with ambient light intensity  $\sim 150\text{-}200$  lx, and without any direct light source above the table. The surface's table was visually divided across both horizontal and vertical axes, resulting in four identical rectangular sectors to account for near/far tablet distances and non-dominant/dominant tablet positions. When required, the users sat on a chair 47 cm high. The marker detection application was running on a Samsung Galaxy Note 10.1 tablet using the computer vision algorithms provided by Vuforia<sup>TM</sup> for Android devices. Visual feedback about the position of the marker with respect to the camera was given by video on the display. The size of this video region was 1/9 the screen size, and it was located in the lower-left corner. The size of the big and small markers was  $50 \times 50$  mm and  $17 \times 17$  mm respectively, and both were printed on a cardboard card of dimensions  $63 \times 91$  mm and situated near the top (see Fig. 2).

### 3.2 Participants

Thirty-two volunteers, 16 male and 16 female, participated in this study. Ages ranged from 24 to 48 ( $M = 32.28$ ,  $SD = 6.3$ ). The average height of males was 180.88 cm ( $SD = 7.27$ ), and females' was 167.38 cm ( $SD = 7.14$ ). Four of them were left-handed and the rest, right-handed.

### 3.3 Design

Since the number of factors considered is high and it would be impractical to have each user perform every combination of their levels, the experiment follows a mixed fractional factorial design  $2^{7-3}_{IV}$ , for which only 16 different treatments are needed; and,

since gender is a between-subjects factor, each user only had to conduct 8 treatments. A total of 1536 interactions were performed across factors and treatments. To avoid order and carryover effects during the performance of the 8 treatments by a given subject, the order in which the treatments were presented follows an  $8 \times 8$  balanced Latin squares design for each gender.

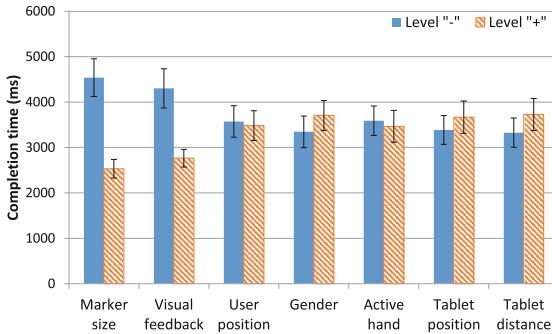
### 3.4 Procedure

Firstly, the users were given some time to train with the system in order to minimize posterior learning effects. During this training, they practiced the experimental task which consisted of taking the card with the marker and bringing it closer to the tablet's camera for it to be recognized. Once they felt familiar with the interaction, the proper experiment began, where one user at a time had to repeat the previous interaction but following the instructions given by the different factor treatments (i.e., being seated or standing, holding the marker with their dominant or non-dominant hand, having the tablet near or far, etc.). Each treatment was repeated six times, and subjects were encouraged to perform this interaction as quickly as possible. For each treatment, the average elapsed time to the detection of the marker was measured, and the users filled a NASA-RTlx questionnaire [3] to assess subjective task load. Once all treatments were complete, a System Usability Scale (SUS) questionnaire [2] was administered to evaluate the usability of the technique. Users were also asked about their experiences in a short post-task interview.

### 3.5 Results

Regarding quantitative differences, Fig. 3 depicts the mean completion times by each factor. A repeated measures ANOVA (with an  $\alpha = 0.05$ ) revealed only two factors having a significant effect on the response variable: the marker size ( $F_{1,243} = 19.514$ ,  $p < 0.001$ ) and the visual feedback ( $F_{1,243} = 11.555$ ,  $p = 0.001$ ). No double or triple interactions were found significant. The shorter times correspond to performing the interaction with the big marker and having visual feedback, respectively.

NASA-RTlx scores for workload were analyzed using Friedman's  $\chi^2$  tests for ranks ( $df = 1$ ). Subjects reported relatively low levels of mental, physical, and temporal demand for all conditions, as well as low levels of effort and frustration. All of the above were rated between 20 and 30 approx., with 0 meaning "very low", and 100 "very high". However, some aspects show significant differences between the levels of factors in some conditions. Participants reported a significant lower degree of workload in general using the big marker and having the tablet near. Concretely, these two conditions received significantly ( $p < 0.05$ ) lower scores of mental/temporal demand, effort, and frustration. Subjects also made less effort ( $p = 0.034$ ) when the visual feedback via video was shown. No significant differences were found between postures, but subjects perceived using their dominant hand as less physical ( $p = 0.009$ ) and time demanding ( $p = 0.009$ ). As for differences between gender, women reported significantly ( $p < 0.05$ ) lower levels of workload in general than men. Concretely, they reported lower levels of mental demand, effort, and frustration, and they showed more confidence regarding performance than men, whose scores were more neutral.



**Fig. 3.** Mean completion times (in milliseconds) for each considered factor

The perceived general usability of the ADI technique was studied via a SUS questionnaire. The analysis using Mann-Whitney U tests revealed no significant differences between genders ( $U = 90.5, p = 0.156$ ). The SUS total score (calculated as it is explained in [2]) is, on average, 73.13 ( $SD = 11.46$ ) for men, and 75.63 ( $SD = 20.18$ ) for women, which, according to Sauro's guidelines [12] is above average (68). In Sauro's letter-grade system (from A + to F), the overall usability of our technique would receive a B.

### 3.6 Discussion

The results show that the proposed interaction technique is usable in general (obtaining a B grade from the SUS questionnaires). This was also confirmed by some subjects' comments regarding the good performance of the system and their suspicions about whether this was a real working prototype.

Further, the quantitative analysis has not found any statistically significant effects related to the posture of the users (sitting vs. standing), the side at which the device is located (dominant vs. non-dominant) and its distance from the user (near vs. far). This proves that the technique is usable under a wide spectrum of ergonomic situations. Moreover, the fact that subjects reported low levels of mental, physical, and temporal demand for all conditions means that the analyzed interaction is intuitive and potentially applicable to more demanding scenarios with subjects having cognitive or motor disabilities. However, the analysis of the qualitative results shows some implications for future applications using the interaction technique considered in this work. Several participants reported having difficulties with the visibility of the video feedback when they were seated and the tablet was not located within arm's reach. This issue disappeared when they were standing up. Hence, this suggests that activities designed for situations where the users are seated should either keep the devices within arm's reach for all participants or avoid video feedback.

Subjects also reported using their dominant hand was less physically and time demanding. However, this did not have a significant impact on the time to perform the gesture and some participants even commented they preferred to use one hand or the other depending on which side the tablet was on. This observation provides initial clues

about the feasible use of both hands to allow bi-manual aerial interactions in the design of future ADI systems. Visual feedback and marker size had a greater impact on the interaction. The small marker was harder to recognize by the application, and the users needed to put it very close to the camera, which made them “lean too much” on the table when the tablet was distant. In a real application, this could cause interference with others’ actions and lead to much more frustration. Therefore, the results suggest that big markers are more suited for this sort of interaction. They can be attached to cards and, because the proper marker is a frame, they can be filled with application-specific content. Taking advantage of markers’ IDs, a potential use for these cards could be using them as information containers and also for the transfer of elements/documents in a co-located group, which many subjects of our experiment showed great excitement about. According to Nacenta et al. [10], an interaction like this one, using the card as an information container, presents intrinsic benefits in terms of awareness since the system does not need to present any action points to the user (e.g., via a cursor), and because any user can easily identify which colleague performed a given gesture with the card because of the visibility of such actions. On the other hand, having visual feedback in a small region of the screen was perceived as a very useful feature for allowing users to adjust and correct their actions. Similar benefits from visual feedback are reported by Nacenta et al. [10]. Nevertheless, some subjects pointed out that it might not be convenient to integrate this feature, since it would reduce the display work area.

## 4 Limitations and Future Work

It is important to note that, as Nacenta et al. [10] remark, the choice of a given interaction technique affects coordination, preference, and performance. Hence, the results reported in this paper are only conclusive for this particular interaction. There are also several limitations to our work. Firstly, the experiments were performed with the same tablet brand and model, and the resolution of the camera could have an effect on the recognition. Secondly, the study only refers to the initial acquisition phase of the interaction. However, analyzing the feasibility of this initial phase and obtaining first user impressions is a necessary step forward before designing and considering more complex scenarios based on ADI. Regarding this, the results obtained provide useful information for future designers of ADI-based environments using frame markers. As a next step, we plan to perform a full study of different types of ADI marker-based manipulations to obtain a full set of feasible interactions with this technique. Another limitation is that the controlled interaction was performed in isolation by a single user and, therefore, no interference issues that could happen in a collaborative scenario were evaluated. This is certainly an interesting area of future research to evaluate the full potential of the proposed technology to support MDE collaboration.

Despite these limitations, the relatively good usability results obtained encourage us to delve into the ADI proposed in this paper, and conduct further experiments that test its suitability in actual collaborative environments. Regarding future uses, we consider this interaction to be promising in meeting environments where users could easily exchange documents attached to marker cards. Another potential application domain is gaming, where, for instance, the tablets could be used in consonance with a physical

gaming board offering digital augmentation, whereas the cards could encode abilities or objects to be transferred to the other participants. Physical games could also be implemented in which participants would exercise by moving around a big table with their cards obtaining and depositing items from the tablets scattered on it. This ADI could also be interesting in rehabilitation tasks for people with acquired brain injuries. In this case, markers could be attached to tangible elements that represent objects in real life and the patients should have to reach the tablet (among several others) that shows some digital content related to the element they are holding.

**Acknowledgements.** This work received financial support from Spanish MINECO (projects TIN2010-20488 and TIN2014-60077-R), from Universitat Politècnica de València (UPV-FE-2014-24), and from GVA (APOSTD/2013/013 and ACIF/2014/214).

## References

1. Avrahami, D., Wobbrock, J.O., Izadi, S.: Portico: tangible interaction on and around a tablet. In: UIST 2011, pp. 347–356. ACM (2011)
2. Brooke, J.: Sus: a quick and dirty usability scale. In: Jordan, P.W., et al. (eds.) Usability Evaluation in Industry, pp. 189–194. Taylor and Francis, London (1996)
3. Byers, J.C., Bittner, Jr., A.C., Hill, S.G.: Traditional and raw task load index (tlx) correlations: Are paired comparisons necessary? In: Mital, A. (ed.) Advances in Industrial Ergonomics and Safety, pp. 481–485. Taylor and Francis, London (1989)
4. Gutwin, C., Subramanian, S., Pinelle, D.: Designing digital tables for highly integrated collaboration. Technical report HCI-TR-06-02, University of Saskatchewan (2006)
5. Hasan, K., Ahlström, D., Irani, P.: Ad-binning: leveraging around device space for storing, browsing and retrieving mobile device content. In: CHI 2013, pp. 899–908. ACM (2013)
6. Hornecker, E., Marshall, P., Dalton, N.S., Rogers, Y.: Collaboration and interference: awareness with mice or touch input. In: CSCW 2008, pp. 167–176. ACM (2008)
7. Katabdar, H., Yüksel, K. A., Roshandel, M.: Magitact: interaction with mobile devices based on compass (magnetic) sensor. In: IUI 2010, pp. 413–414. ACM (2010)
8. Kratz, S., Rohs, M.: Overflow: expanding the design space of around-device interaction. In: MobileHCI 2009, pp. 4:1–4:8. ACM (2009)
9. Marquardt, N., Kiemer, J., Greenberg, S.: What caused that touch?: expressive interaction with a surface through fiduciary-tagged gloves. In: ITS 2010, pp. 139–142. ACM (2010)
10. Nacenta, M.A., Pinelle, D., Stuckel, D., Gutwin, C.: The effects of interaction technique on coordination in tabletop groupware. In: GI 2007, pp. 191–198. ACM (2007)
11. Rick, J., Marshall, P., Yuill, N.: Beyond one-size-fits-all: how interactive tabletops support collaborative learning. In: IDC 2011, pp. 109–117. ACM (2011)
12. Sauro, J.: Measuring usability with the system usability scale (sus), February 2011. <http://www.measuringusability.com/sus.php>. Accessed March 2015
13. Tuddenham, P., Kirk, D., Izadi, S.: Graspables revisited: multi-touch vs. tangible input for tabletop displays in acquisition and manipulation tasks. In: CHI 2010, pp. 2223–2232. ACM (2010)

# On Applying Experience Sampling Method to A/B Testing of Mobile Applications: A Case Study

Myunghee Lee and Gerard J. Kim<sup>(✉)</sup>

Digital Experience Laboratory, Korea University, Seoul, Korea  
`{revisel,gjkim}@korea.ac.kr`

**Abstract.** With the advent of mobile devices, the experience sampling method (ESM) is increasingly used as a convenient and effective way to capture user behaviors of, and evaluate mobile and environment-context dependent applications. Like any field based in situ testing methods, ESM is prone to biases from unreliable and unbalanced data, especially for A/B testing situations. Mitigating such effects can in turn incur significant costs in terms of the number of participants and sessions, and prolonged experimental time. In fact, ESM has rarely been applied to A/B testing nor do existing literatures reveal its operational details and difficulties. In this paper, as a step toward establishing concrete guidelines, we describe a case study of applying ESM to evaluating two competing interfaces for a mobile application. Based on the gathered data and direct interviews with the participants, we highlight the difficulties experienced and lessons learned. In addition, we make a proposal for a new ESM in which the experimental parameters are dynamically reconfigured based on the intermediate experimental results to overcome the aforementioned difficulties.

**Keywords:** Experience sampling method (ESM) · A/B testing · Usability

## 1 Introduction

Experience (or environment) sampling method (ESM) is a system evaluation and behavior capture method by which user evaluative responses are made and recorded at the exact time and place of the system usage. Compared to the old paper-and-pencil method, with the advent of mobile devices, ESM can be carried out more conveniently e.g. as or through functionalities embedded in smart phone sensors and applications. Like any field-based in situ testing methods, ESM can suffer from biases that might otherwise be controllable in a laboratory setting, but at the same time, they can be mitigated through a high number of repetitions, extended length of experimentation, a large number of participants, and thus at higher cost. However, this can also ironically bring about even more unreliable and unbalanced data. This is more problematic in the case of a comparative evaluation in which, for a validity and fairness, it is necessary to collect a minimum and “balanced” amount of reliable data. In fact, ESM has rarely been applied to A/B testing nor do existing literatures reveal its operational details and difficulties. In

this paper, as a step toward establishing concrete guidelines for A/B testing with ESM, we describe a case study of applying ESM to evaluating two competing interfaces for a mobile application. Based on the gathered data and direct interviews with the participants, we highlight the difficulties experienced and lessons learned. In addition we make a proposal for a new ESM in which the experimental parameters are dynamically reconfigured based on the intermediate experimental results to overcome the aforementioned difficulties and run the experiment more economically.

## 2 Related Work

The Experience Sampling Method (ESM) was first introduced by Larson and Csikszentmihalyi as a research tool in social science [1], but has found great utility especially in mobile HCI research [2]. For instance, the Context Aware Experience Sampling Tool developed by Intille et al. [3], one of the first of its kind, allowed a flexible data solicitation by scripting in the survey questions and multiple choice answers, and including a functionality for users to easily capture and store multi-media data on a PDA. Consolvo et al. used a similarly designed ESM tool, called the iESP, to evaluate ubiquitous computing applications and further analyzed the possible pitfalls and lessons learned of using such a methodology for in situ HCI evaluation (e.g. the effectiveness of self-reporting using the mobile devices and the need for tailoring the data collection process for the target subjects) [4]. In 2007, Froehlich et al. also introduced a more advanced mobile based ESM tool called MyExperience which offered an XML based specification method of how to solicit data from the user, sensor based context triggered sampling and structured storage of logged data to a data base server [5]. Momento, a tool developed by Carter et al. is another step in the evolution of the mobile based ESM tools offereing sampling control and on-line monitoring (e.g. visualization and analysis of incoming data) from a remote desktop server [6]. Maestro further extended the sampling control capability by exploiting long term user behavior and usage patterns for shaping personalized ESM questions to different types of users [7]. As ESM tools become more refined and enter into one of the main stream evaluation methods, its user interface/interaction design itself has emerged as an important issue as well with regards to the requirement and desire to encourage the participants to make faithful and reliable response [8]. While not usually employed in ESM, the data collection in crowdsourcing can involve “gold standard” questions to ensure the reliability and credibility of the contributors [9]. Answering performanes to the gold standard questions can be used to exclude certain data, e.g. those that are regarded too mechanical or even those of programmed bots.

In summary, it can be seen that ESM tools are continuing to evolve and being added with more functionalities (for both the participants and experiment administrators) and the methodology extended for a more reliable and credible results. In most previous related work we have reviewed, ESM was still used for capturing context dependent user behaviors for a “single” application. According to the recent survey of ESM tools by Conner [2], the data collection schedule and design are fixed

throughout the experiment, e.g. choice of participants, number of participants, sampling time, experiment duration, etc. ESM tool capabilities and methodological process need to be further extended for scalability and efficiency to handle larger subject pools, longitudinal studies and A/B testing with several factors.

### 3 Case Study: ESM for A/B Testing

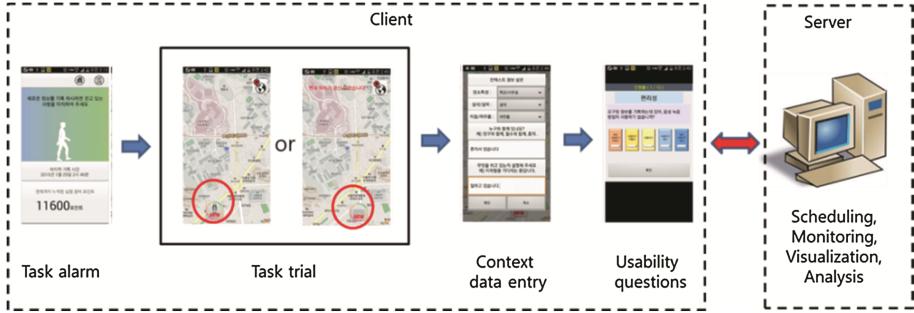
#### 3.1 Test Application and Evaluated Interfaces

In this case study, we evaluate the usability of two competing interfaces for a simple mobile map-logging application (see Fig. 1) in which a user can record and tag short information about the current user location (indicated on the map and sensed by the GPS sensor). The two competing interfaces compared were for entering information through (1) voice and (2) touch typing of text. The ESM is used because the application is mobile and possibly context dependent (e.g. location, time, social setting, etc.). Either by voluntary initiation or by a scheduled prompt, the participant is asked to try out one of the interfaces (chosen in a balanced order) to enter short information (i.e. record in voice, or type in). In addition, several supplemental contextual information (that cannot be easily inferred automatically with the sensors) is solicited using a menu driven interface, asking whether the participant is indoor or outdoor, the location (e.g. restaurant, classroom, streets), on-going activity (e.g. resting, eating, in a meeting), social situation (e.g. alone, with a friend), etc.

#### 3.2 ESM Based A/B Testing Process and the Support Tool

Typically, a comparative UI experiment is conducted in a laboratory setting as a cross sectional study with repeated measures taken in batch. On the other hand, ESM is used to attain more relevant data considering the environment and context of usage, but batch collection of repeated usage is often not feasible. Rather the data is collected over some extended period of time. Nevertheless, we still regard our experiment to be “cross-sectional” since we are not (for now) interested in longitudinal change in the user response. Note ESM has primarily been used for capturing user behavior for a single application rather than for making comparisons of usability or UX. In our case, the ESM A/B testing proceeds as a within-subject type (i.e. the user tries out both interfaces and make comparisons) for a given period of time with a predetermined number of participants to first collect some minimum amount of data deemed sufficient for the power of the experiment. While there are several methods to decide on the least amount of required data or number of participants, for now the experiment duration, amount of data solicited (equally number of repetition) and the number of participants were determined arbitrarily but in a conservative fashion (e.g. long enough to gather good amount of data).

The participants were scheduled in a balanced order to fulfill a task, either using the interface A or B, and make answers to usability and gold standard questions. Due to the difficulty in inferring particular usage contexts automatically (e.g. whether a person is moving, in a meeting, at a bus station, sleeping, etc.), the data solicitation



**Fig. 1.** The overall process of ESM based A/B testing of two interfaces (using the voice or text input) for a simple mobile map logging application.

was done according to a regular time schedule rather than invoked by automatic context detection. Figure 1 shows the overall flow of the ESM A/B testing process.

### 3.3 Detailed Experiment Procedure

The first phase of the ESM based A/B testing was conducted for six days, and solicited for data entry 8 times a day. For now, the duration of one week for the first phase was decided rather arbitrarily. A total of 30 subjects (20 males/10 females) mostly in their 20's with various occupational backgrounds participated in the study. The participants were recruited, interviewed and selected through an on-line social messaging system. The participants were given instructions as how to download and install the smartphone application, and how and when to try out the tasks and make proper data entries. Prior to the actual experiment, the participants were given instructions for a short training session for getting oneself familiarized with the application, two interfaces and data entry method.

The participants were compensated upon the completion of the whole session at rate of \$0.23 per answered questions (which totaled to about \$11 dollars maximum).

Considering the recommendation by [10], the task trials and data solicitations were scheduled every two hours only between 7:30 am to 10:30 pm (total of eight times a day). At the scheduled times, the application was invoked on the smartphone device automatically (4 times each for the respective interfaces in an alternative order), and an alarm was used to remind and notify the participant. Despite the reminders, it was up to the participants to actually respond. It was also possible that the alarm or smartphone itself was switched off. Thus, a few simple user behavioral checking measures were implemented. For example, 30 s of no response was regarded as a refusal of a data entry. It was also checked whether the phone was actually in active use before and after the scheduled time to guess whether the “refusal” was deliberate or not. Such piece of behavioral information was to be used collectively to assess the credibility or reliability of the participant. Upon a scheduled invocation of the task trial, the user was to enter information as attached to wherever the user was at, enter additional contextual information (as described in Sect. 3.1) and answer a series of usability questions in a 5 Likert scale (on convenience and ease of use,

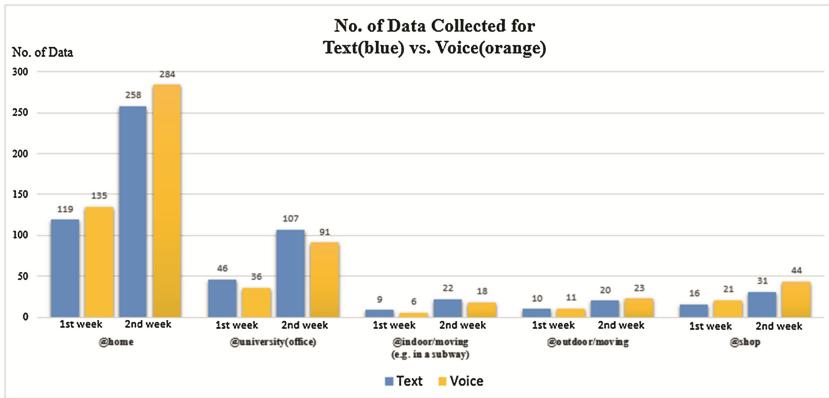
general satisfaction, annoyance, fatigue). Finally, gold standard questions were given to assess and explicitly confirm the credibility and reliability of the participant. The gold standard questions were designed to be fairly easy with the least cognitive burden, yet not answerable by random guesses, such as solving a simple arithmetic (e.g. “what is  $(5 + 4) * (1 + 1)$ ?”) or asking of common sense knowledge in multiple choices (e.g. “who is the president of Korea?”). After completing the task of logging the map either by voice or text, there were a total of 10 questions to answer including the two gold standard quiz. After six days, the collected data were analyzed for sufficiency and (as explained more in detail in the next section) it was determined that another round of data collection was deemed necessary. A second phase of data collection for A/B testing was continued for another week. After the whole two week sessions, we administered one final survey, asking the participants about the ESM procedure itself (participants were separately compensated for it). The answers were used, in addition to the operational problems discovered during the case study, to base our proposal for an improved and extended ESM. We omit the presentation of the detailed survey questions and only brief the results in Sect. 4.4.

## 4 Results, Observations and Proposals

### 4.1 First Phase: Data Sufficiency and Balance

During the first phase (first six days), with 30 participants and 8 data collection sessions per day, we ideally would have collected 1440 sets of balanced and reliable task trial and session response data. However, only 463 session data (37.2 %) were collected due to reasons such as participant not noticing the incoming alarm (41.8 %) and deliberate refusals (20.3 %). At a closer look, the collected data were even more insufficient with respect to different contexts and number of participants. For instance, the data collected for @home context comprised more than half of the total data, while the rest scattered in little proportions to other usage location contexts and thus lacking the power for any meaningful analysis. The situation was worse for conjunctive contexts such as for a particular location and time, location and activity, etc. This was not only attributed to the fact that the data were collected based on a simple time based schedule (rather than based on intelligent, but technically difficult, context detection), certain context based usages just do not happen as often as others (e.g. staying home vs. riding on a subway). It was also possible that by the nature of the application, the users just was not up to using the application as often as necessary to gather sufficient interaction data in a short amount of time.

Thus, in order to reduce the experiment duration, save cost, relieve the burden on the participants and ultimately make the study more focused rather than open-ended and exploratory, we propose that intermittent data analysis would be necessary (as part of an extended ESM A/B testing study methodology) to check data sufficiency and analysis power, carry out the mid-evaluation if possible, and eliminate certain dependent variable measurements if the analysis results are clear (e.g. very low p-value, high  $R^2$ , high  $\chi^2$  etc.). In this study, the experiment continued on for another week (second phase) and about the same amount of data were additionally collected (See Fig. 2). The comparative



**Fig. 2.** Comparative usability data collected during the first six days (first two bars among the four) and after the second week (cumulative, the second two bars) for different location contexts (@home, @university, etc.). The dark blue bar represents data for text, and the light yellow for voice based interfaces (Color figure online).

usability between the text and voice based interfaces for the context of @home usage (which had sufficient data for analysis after the first week) did not change. Note that comparative qualitative assessments of the interfaces can be gathered as well but would require a subjective analysis.

In addition, the data collection process and scheduling must be tailored toward the particular context of interest. If automatic context detection is technically difficult, then a pilot study should be conducted ahead of time e.g. to recruit a participant who is likely to make a particular context-based usage of interest, or personalize the data entry session schedule at a right timing by an analysis of one's daily activities. Furthermore, the data were unbalanced in many ways, e.g. between the treatments (e.g. data for voice based interface vs. text based), among the participants, and among the contexts. The missed data entry sessions, deliberate or not, did not originate uniformly among different participants. While the unbalanced amount of data between treatments is a serious problem to making comparative analysis, it can be viewed as an indication of usability or preference. Thus, through continuous monitoring, the ESM must be administered in such a way to solicit data and closely balance the competing data e.g. by scheduling for treatments that lack data, encouraging the non-responding participants, and analyzing whether the unbalanced response is in fact due to preference or certain operational constraint or contexts. Note that such provisions can contaminate the balanced presentation of treatments (for mitigating the learning effect), thus must be applied carefully in an incremental manner.

## 4.2 Participant Reliability

The gold standard quiz is only a partial and indirect indicator of participant/data reliability. At any rate, we judged that the data themselves were reasonably reliable

and credible because only less than 5 % of the gold standard quizzes were incorrect overall. In addition, the response behavior did not change much over the second week. Thus, it seemed more important to single out the participants who tended to “refuse” (especially deliberately so) the data entry in the first place too often. Future ESM tools should have the support capability for monitoring for these participants and replacing them if necessary.

#### 4.3 Experiment Extension and the Utility of ESM

Because sufficient data were not collected for a meaningful comparative analysis for different contexts of usage (except for the @home usage), the experiment was continued on for the second week (see Fig. 2). Future ESM tools should administer such an extended experiment in a systematic fashion through data analysis. Note that during the second week, the missed data entry session and data unbalance were still at the similar level. Future ESM tools must take measures to minimize these types of data insufficiency. On the other hand, sufficient data were then collected for the @university usage case, and showed different usability results from the @home usage (e.g. participants preferred text based input more for @university than @home). Thus, this at least confirms the very utility of the ESM in that it can capture different usability and user behavior depending on the usage context. After the additional experimentation, nothing much has changed except only the data for @university usage context became sufficient (for power of analysis). Data for other contexts were still lacking and unbalanced, and the additional data for @home usage did not change the initial analysis results.

#### 4.4 Participant Responses About the ESM Process Itself

Participants mostly acknowledged the experimenter’s sentiment of the difficulty in collecting reliable data. They suggested for a system of incentive based compensation, better alarm mechanisms, and pre-notification of the upcoming data entry sessions. Two main reasons for the missed data entry were not being able to notice the alarms and scheduled events overlapping with uninterrupted on-going activities. They also expressed that the gold standard questioning was effective not only as an indicator for credibility but also encouraged the participants to be more thoughtful and reliable.

### 5 Conclusion and Future Work

In this paper, we described a detailed case study of applying ESM to evaluating two competing interfaces for a mobile application. Based on the gathered data and direct interviews with the participants, we highlighted the difficulties experienced and lessons learned. In addition, we made several proposals for a new ESM (also our future work) with the capabilities to flexibly revise the parameters of the experiment on-line so that the ESM can be run economically, efficiently but with the same reliability.

**Acknowledgements.** This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (No. 2011-0030079) and funded by the Forensic Research Program of the National Forensic Service (NFS), Ministry of Government Administration and Home Affairs, Korea. (NFS-2015-DIGITAL-04).

## References

1. Larson, R., Csikszentmihalyi, M.: The experience sampling method. *New Dir. Methodol. Soc. Behav. Sci.* **15**, 41–56 (1983)
2. Conner, T.S.: Experience sampling and ecological momentary assessment with mobile phones (2013). <http://www.otago.ac.nz/psychology/otago047475.pdf>
3. Intille, S.S., Rondoni, J., Kukla, C., Ancona, I., Bao, L.: A context-aware experience sampling tool. In: Proceedings of SIGCHI Conference Extended Abstracts, pp. 972–973. ACM (2003)
4. Consolvo, S., Walker, M.: Using the experience sampling method to evaluate ubicomp applications. *IEEE Pervasive Comput.* **2**(2), 24–31 (2003)
5. Froehlich, J., Chen, M.Y., Consolvo, S., Harrison, B., Landay, J.A.: MyExperience: a system for in situ tracing and capturing of user feedback on mobile phones. In: Proceedings of the International Conference on Mobile Systems, Applications and Services, pp. 57–70. ACM (2007)
6. Carter, S., Mankoff, J., Heer, J.: Momento: support for situated ubicomp experimentation. In: Proceedings of the SIGCHI Conference, pp. 125–134. ACM (2007)
7. Meschtscherjakov, A., Reitberger, W., Tscheligi, M.: MAESTRO: orchestrating user behavior driven and context triggered experience sampling. In: Proceedings of International Conference on Methods and Techniques in Behavioral Research, p. 29. ACM (2010)
8. Consolvo, S., Harrison, B., Smith, I., Chen, M.Y., Everitt, K., Froehlich, J., Landay, J.A.: Conducting in situ evaluations for and with ubiquitous computing technologies. *Int. J. Hum. Comput. Interact.* **22**(1), 107–122 (2007)
9. 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 SIGCHI Conference, pp. 2399–2402. ACM (2010)
10. Abdesslem, F.B., Parris, I., Henderson, T.N.H.: Mobile experience sampling: reaching the parts of Facebook other methods cannot reach. In: Proceedings of Privacy and Usability Methods Powwow (2010)

# Usability Aspects of the Inside-in Approach for Ancillary Search Tasks on the Web

Marco Winckler<sup>2(✉)</sup>, Ricardo Cava<sup>1</sup>, Eric Barboni<sup>2</sup>,  
Philippe Palanque<sup>2</sup>, and Carla Freitas<sup>1</sup>

<sup>1</sup> Institute of Informatics, Federal University of Rio Grande do Sul,  
Av. Bento Gonçalves 9600, PB 15064, Porto Alegre, RS 91501-970, Brazil  
[{racava, carla}@inf.ufrgs.br](mailto:{racava, carla}@inf.ufrgs.br)

<sup>2</sup> ICS-IRIT, University of Toulouse 3, 118, route de Narbonne,  
31062 Toulouse Cedex 9, France  
[{winckler, barboni, palanque}@irit.fr](mailto:{winckler, barboni, palanque}@irit.fr)

**Abstract.** Given the huge amount of data available over the Web nowadays, search engines become essential tools helping users to find the information they are looking for. Nonetheless, search engines often return large sets of results which must be filtered by the users to find the suitable information items. However, in many cases, filtering is not enough, as the results returned by the engine require users to perform a secondary search to complement the current information thus featuring ancillary search tasks. Such ancillary search tasks create a nested context for user tasks that increases the articulatory distance between the users and their ultimate goal. In this paper, we analyze the interplay between such ancillary searches and other primary search tasks on the Web. Moreover, we describe the inside-in approach, which aims at reducing the articulatory distance between interleaved tasks by allowing users to perform ancillary search tasks without losing the context. The inside-in approach is illustrated by means of a case study based on ancillary searches of coauthors in a digital library, using an information visualization technique.

**Keywords:** Interaction gulfs · Web search · Ancillary queries · Nested user tasks

## 1 Introduction

According to Hilbert and López [8], in 2007 almost 94 % of our memory was already in digital form and most of it can be found through the Web nowadays. Over the years, users have become used to retrieve information from the Web, and for that they developed several strategies, which can be summarized as information lookup and exploratory search [9]. Whilst exploratory search requires time for scanning and reading documents, information lookup can be solved by simple factual question-answer interactions. Moreover, in this context of huge amount of data, information retrieval systems and search engines have become an integral part of our daily lives [7].

The user interface provided by such information retrieval systems must be simple enough to allow users to formulate queries and understand the results provided by search engines [10, 18]. Nonetheless, many users are still struggling to use them to obtain the results they need [6]. Many of the problems users have to face are related to increasing availability of data in the Web. For that, users must be very precise in the way they formulate their queries, and they must know how to interact with the display to identify the sought results in the large set of data.

Quite often, queries start by filling a search box with keywords. Formulating queries in this way is a daunting task that requires users to open a new window, to fill in a form with appropriate keywords, and then scan the list of results until finding the one that corresponds to their goal. Moreover, the standard way to display search results (obtained by either filling in a form or browsing documents) often imply to display in a new window/tab and/or replace the current window/tab's contents, which might deviate the users' focus of attention and creates an interruption between nested tasks.

As we will see, while some search tasks can be directly associated to a user's primary goal, many other searches are nested in other tasks and simply require to complement information they are currently reading [7]. For example, users reading an article in a Web page might be curious to know with whom the author of that particular article has published in the past. In this scenario, looking up for co-authors constitutes an ancillary search, which is not meant to divert users' attention from reading the article. For such kind of ancillary-search tasks, displaying results in a new window/tab might be unsuitable since it creates an articulatory distance between the origin of the request and the information display, making difficult to users to assess if their goal has been fulfilled or not by the query.

In this paper we claim that whilst the design of existing tools might fit for the purposes of primary search tasks, users still need better support for performing ancillary search tasks. This claim is supported by a model-based task analysis presented in Sect. 2. Based on the Don Norman's cognitive model [14], we discuss in which extension the design alternatives for performing ancillary search tasks might increase cognitive gulfs of execution and gulfs of evaluation. Based on the lessons learned from the tasks analysis we have devised an alternative approach, called *inside-in* search, which aims at reducing the articulatory distance between interleaved tasks by allowing users to perform ancillary search tasks without losing the context. The *inside-in* approach is presented in the Sect. 3. In order to demonstrate the feasibility of the *inside-in* approach, we have developed a supporting tool that is described in Sect. 4. The last sections of the paper present related work, the conclusions and future work.

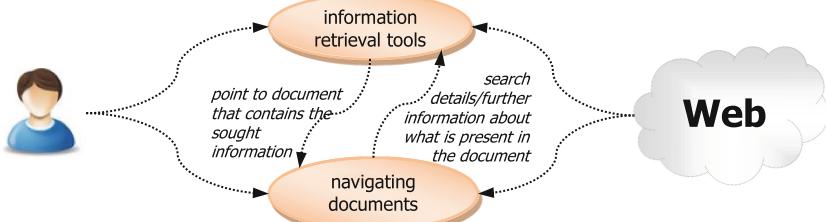
## 2 Task Analysis of Web Search Tasks

In order to support the analysis of users' tasks we employ the model-based notation HAMSTERS, which stands for Human-centered Assessment and Modeling to Support Task Engineering for Resilient Systems [11]. Similar to CTT [15], HAMSTERS provides constructs (i.e., typology of tasks and operators) that allow fine-grained description of tasks as well as the simulation of models. Moreover, thanks to appropriate tool support, HAMSTERS models can be used to predict users' performance

with interactive systems, which can be understood as an indirect (or simulated) knowledge about human behavior. Whilst a detailed description of the notation HAMSTERS is out of the scope of this paper, we provided the necessary information for understanding our models. Further details about HAMSTERS can be found elsewhere [11, 12].

## 2.1 Overview of Search Tasks Over the Web

As argued by Yates and Neto [21] users can search information over the Web either by *navigating documents* and/or using specialized *information retrieval tools* (the so-called search engines). As shown in Fig. 1, even if these tasks are distinct, they are interconnected. On one hand the ultimate goal of search engines is to direct users to Web pages that contain the sought information. On the other hand, documents might contain links embedding queries also pointing to search engines, which are aimed at helping users to find complementary information that is not readily available through a simple navigation [5].

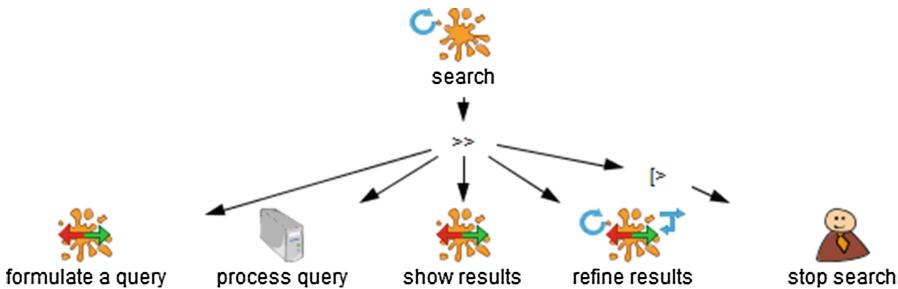


**Fig. 1.** Overview of alternative strategies for finding information; adapted from [21].

The task model described herein aims at analyzing search tasks from a high level of abstraction that do not imply any particular implementation of tools. Indeed, we assume that it would be possible for users to perform searches either by using a dedicated information retrieval tool and/or by triggering a search directly from a Web document. Thus, regardless the information retrieval algorithms and users' needs for information, a search can be summarized as a set of the following subtasks: at first, users *formulate a query*, then the system will *process the query* and *show the results* to the users which can, at this point, to *refine the results* until selecting the appropriate entry that corresponds to the sought information. Moreover, users can decide for any reason, and at any time, to *stop the search*.

One of the main advantages of task model notations such as HAMSTERS is to support the iterative decomposition of tasks in a tree-like structure to reach the level of detail that is required for the analysis. The corresponding modeling of search task using the notation HAMSTERS is illustrated in Fig. 2. Notice that the top-level task *search* is decorated at its left side with the symbol **C**, to indicate that *search* is an iterative task that can be repeated indefinitely by the user. Subtasks are connected by the

operator  $>>$  that determines the sequence of task execution. The operator  $[>]$  associated to the task *stop query* indicates that, when this task is performed, it interrupts the sequence of other subtasks. Notice that specific icons are used to indicate the typology of tasks; for example, *process a query* is typically automated by the system, *stop search* is typically a user task, which might require a simple user decision, and tasks such as *formulate a query*, *show results* and *refine results* require user interaction with the system to be performed, so they are also called interactive. The symbol  $\text{t}$ , next to the task *refine results*, is used to indicate that this task is optional.



**Fig. 2.** Overview of a Web search task described using the notation HAMSTERS.

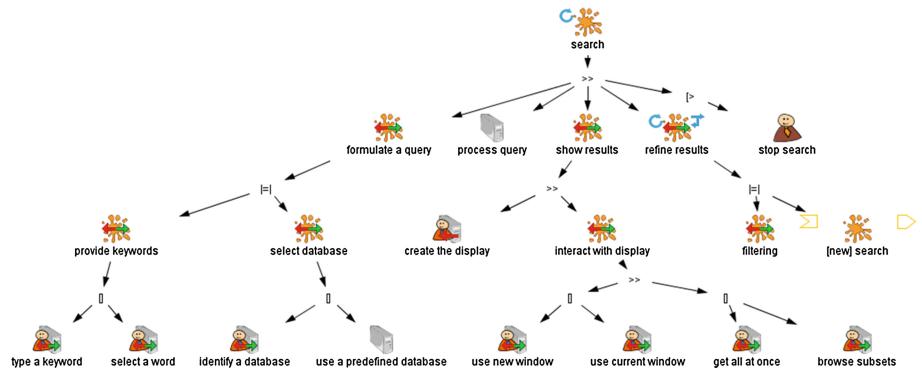
Figure 3 shows a further level of task decomposition with alternative ways to perform some tasks. As for the task *formulate query*, we can identify additional sub-tasks including *provide keywords* and *select a database*. The operator  $| = |$  indicates that these tasks can be performed in any order. It is worthy to notice that the task *provide keywords* is decomposed in alternative subtasks: *type keyword* and *select word*. The choice between tasks is indicated by the operator  $[]$ . Some of the alternative tasks consider the case of manual versus automated execution. For example, the tasks *select database* can be done by prompting the user to explicitly *identify a database*. The automated alternative task assumes that the system *uses a predefined database* and the user cannot change it.

The next task is to *create the display*, which is shown as a simple *output* task in the HAMSTERS' taxonomy. The creation of the display offers several alternatives to perform the task *interact with display*. Basically, from this point, user tasks depend on the location of the display and the number of entries in the set of results. For the location, users might be led to *use new window* or to *use the current window*. For the number of entries, a user can either *get all at once* (i.e. all results appear in the same display) or *browse subset* (i.e. results are divided in subsets that can be navigated by the user).

Users can adopt two main strategies to *refine results*: to apply *filtering* mechanisms or to perform a *(new) search*. A *(new) search* task follows the same pattern of a full *search task*, which means that search tasks can be recursive. To indicate that a subtask corresponds to a pattern of a known task, HAMSTERS provides a particular type of

construct called *copy task*, which is clearly indicated in Fig. 3 as decorations around the tasks (*new search*).

It is noteworthy that only leaf tasks in a HAMSTERS model are ever executed. The level of decomposition of tasks in HAMSTERS is arbitrary, but we assume that the details provided in Fig. 3 are sufficient to illustrate our analysis.

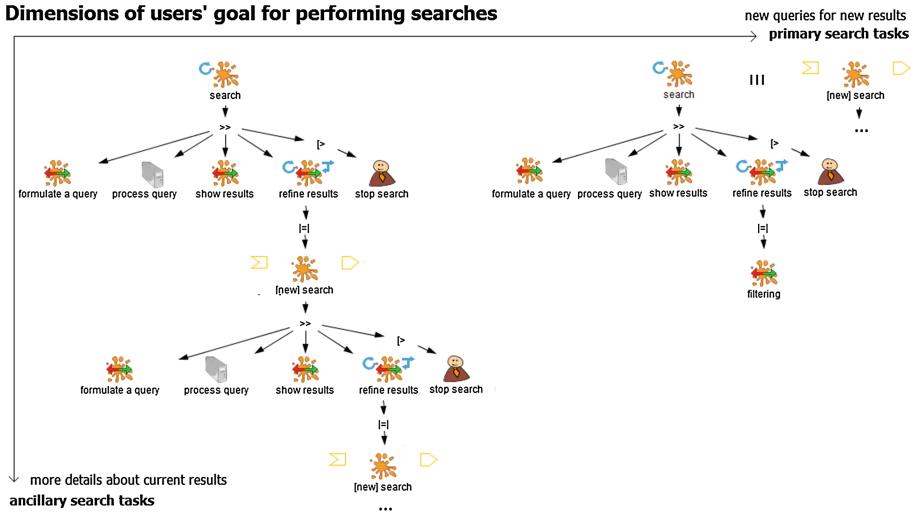


**Fig. 3.** Expanded view of user tasks including alternatives ways for performing a search task.

## 2.2 Overview of Web Search as Primary Task or Ancillary Tasks

By observing users behavior we found that they often have many searches running in parallel (or executed in very short time intervals). Whilst some searches might refer to completely disjoint user's goals (for example, look for a restaurant in town for tonight and plan a trip for the weekend), other searches are indeed part of an overall primary user goal (for example, search for a hotel, then search for a flight whilst planning a trip). It is also interesting to notice that many parallel searches don't correspond to a primary user's goal but they are performed to get information for achieving a previous task (for example, looking for currency exchange rates to calculate prices given in a foreign currency whilst booking a hotel). Figure 4 illustrates the differences between search tasks according to users' goal, which can be formalized as follows:

- *Primary search tasks*, which correspond to a user's primary need for information. Ideally, primary searches encompass a single cycle of question-answer interaction with the search engine. Nonetheless, if the results are not satisfying, users ought to reformulating the terms used in the query and perform a new search. It is interesting to notice that users might perform many queries but every query is treated as unique by the system. As a consequence, the entries in the results provided by different searches might highly differ according to the keywords used.
- *Ancillary search task*, which are aimed at providing details on demand about the results that are current in display. *Ancillary searches* depend on the results obtained from previous search and/or available information over a Web page. Ideally, once users find the answers he should be prompt to return to the context of the task he was performing before launching the *ancillary search*.



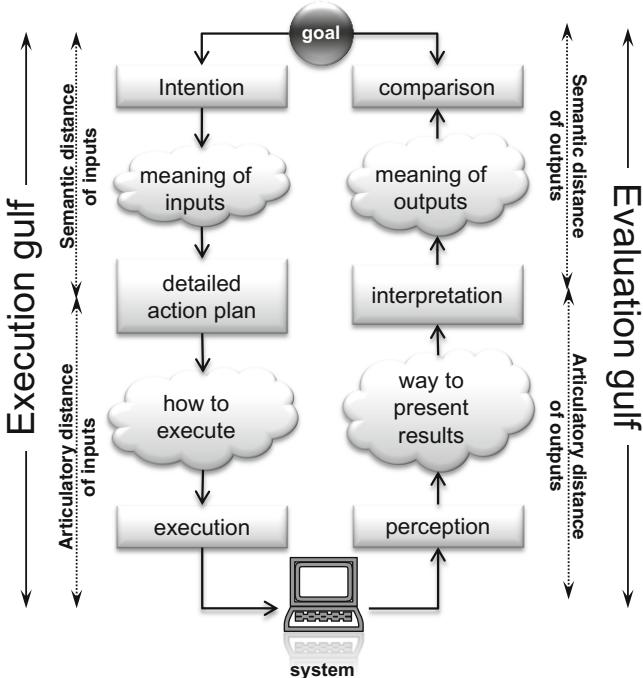
**Fig. 4.** Tasks models in HAMSTERS illustrating dimensions of users' goal for performing search tasks, thus featuring nested *ancillary search tasks* and disjoint *primary search tasks*.

Is worthy of notice that *primary search tasks* are typically treated as disjoint tasks by the system, so they can occur in parallel. However, *ancillary searches* are deeply dependent of existing contents for which it is aimed for providing further details. It is interesting to notice that *ancillary search* leads to nested queries that create a trail of searches performed by the users while *primary search tasks* are treated independently by the system. Both types of tasks can be combined according to the user's actual needs for information. Thus, a *primary search task* can run in parallel with other searches that require the decomposition into *ancillary searches*.

### 2.3 Execution and Evaluation Goals in Web Search Tasks

The tasks that users perform during a search establish a type of question/response communication between the user and the system. Based on Don Norman's cognitive model [14], it is possible to assess the communication mismatch between user's internal goals for performing a search task and the user's expectations with respect to the availability of information specifying the state of the world (the Web). Communication mismatches occur in terms of inputs (i.e., *gulf of execution*), output (i.e. *gulf of evaluation*), or both [13]. In order to illustrate the meaning of the *execution gulf* and the *evaluation gulf* in search tasks, we present in Fig. 5 a revised version of the Norman's cognitive model explicitly showing the *articulatory distance* and the *semantic distance*, both in terms of user *input* (i.e. when users formulate the query) and in terms of system *output* (i.e. when the system shows the results to be assessed by the user).

The length of the *gulf of execution* is described by Norman as the difference between the intentions of the users and what the system allows them to do or how well the system supports those actions. In the case of a Web search, if the users find an



**Fig. 5.** Execution and evaluation gulf in search tasks, adapted from [14].

unknown word while navigating the web, they might expect that clicking on a link (on that word) would provide them with the complementary information required to understand the meaning of the word. In the user's language, "click the link" defines the goal for obtaining the word's meaning. However, if the link does not provide the expected results, users have to execute additional actions, such as opening a new window, visiting a search web site, typing the adequate keywords to specify the search, and, finally, browsing the list of results until getting the desired definition.

The *gulf of evaluation* refers to the way the results provided by the system are meaningful or understandable by the users and in accordance with their goals. In other words, the *gulf of evaluation* is the degree to which the system or artifact provides representations that can be directly perceived and interpreted in terms of the user's expectations and intentions. Thus, if the system does not "present itself" in a way that lets the users derive which sequence of actions will lead to the intended goal, or infer whether previous actions have moved them closer to their goal, there is a large *gulf of evaluation*. In our case, users must spend a considerable amount of effort and significant attentional resources to perform a query in a new window, to identify the answers that correspond to their expectations and, then, to put the word's meaning back in the appropriate context.

Table 1 shows a comparison between the user interface alternatives for tasks (actually, we only take into account the leaf subtasks). A large *execution gulf* is expected when users have to *formulate a query*. Nonetheless, the semantic distance for

**Table 1.** Tradeoffs of design alternatives for performing tasks in a Web search.

Subtasks	User interface alternatives for tasks	Execution gulf		Evaluation gulf	
		Semantic distance of input	Articulatory distance of input	Articulatory distance of output	Semantic distance of output
formulate a query	provide keywords	know how and where to type keywords, know the spelling	require a fill-in-form, typing is time consuming and error prone	the size of the fill-in-form should accommodate the keyword	require recognize what was typed
		select a word	might require a single click	perceive a word as link	recognize the selected word
	select database	identify a database	user must know the database	see the source of information (database) of the items in the results	recognize the source of the results
		use predefined database	no user input required		
show results	interact with display	use new window	predict where results will be shown	move to the new window to see results	require to locate new window in the display
		use current window		keep focus on the display	require to locate results in the current display
		get all at once	users might be prompted to select where they prefer to see results	might require a simple selection	require to locate scroll in the display and eventually use it
		browse subset		might require (or not) users to set the number of entries in the subset	require to locate interactors, navigate between subsets and eventually use it
refine results	filtering	know how to apply filters	select filters	perceive filters and results	recognize results that correspond to filters
	[new] search	require a new instance of search task...			

*Legend:*  short distance,  large distance,  depends on the context.

the task *select word* is smaller than *type keyword* because recognition of words is less cognitive demanding than choosing the words. The *articulatory distance of input* is also shorter for the task *select a word* as it just requires a click, which is also faster than *typing keywords*. Moreover, users can make mistakes (e.g., use the wrong word) and slips (e.g., introduce typos) while typing keywords.

For the *evaluation gulf*, users have to check if the keywords they provide appear in the display properly with no error. The alternatives for *selecting a database*, to *identify*

a database (either by typing the name of the information source or selecting it from a list) has larger *semantic* and *articulatory* distances of input when compared to the *use of a predefined database* as it is known by the system and doesn't require user interaction. For the *evaluation gulf*, the semantic and articulatory distances are similar in both cases as users must be able to recognize information sources whatever it is.

The tasks *interact with display* implies alternatives for the location of the display (*new window* or *current display*) and for the number of entries in the set of results (*get all [results] at once* or *browse subsets*). As far as location is a concern, we might say that all options have similar semantic distance of input, since users must be able to predict where results will be shown. However, *use current window* for showing results is less demanding in terms of articulatory distance of input as users can keep the focus on the display and do not need to move to another window. Using a new window for the display would also require users to locate where results are located in the display, which increases articulatory distance of output. For the semantic distance of output, *use new window* requires to manage multiple windows while *use current window* implies that previous content of the window is lost or new content is superposed to the existing one, which might be confusing. The relative advantage of these options can only be decided once the context for the user search is known.

The options for the task *refine results* are: *filtering* and perform a (*new*) *search*. Many *filtering* mechanisms exist and a deep analysis of them is out of the scope of this paper, but we can assume that filtering might help to locate an item of information in the set of results. When the sought information cannot be found in the set of results in display, the only alternative users have is to make a (*new*) *search*. The execution of a (*new*) *search* is recursive, and requires users to go through all the subtasks. For that reason, we consider that making a (*new*) *search* requires extra effort from users when compared to exploring results that are already in the display.

Table 2 provides an analysis of the tradeoffs between the two alternative strategies for performing a search task. As we can see, users might perform a (*new*) *search* as a *primary search task* or as an *ancillary task*. If the *search task* is performed as a *primary task*, users have to manage the articulation of possibly disjoint searches performed in parallel and formulate new queries from scratch. Assuming that users expect to keep

**Table 2.** Tradeoffs between Web searches as primary and ancillary tasks.

Web search	Execution gulf		Evaluation gulf	
	Semantic distance of input	Articulatory distance of input	Articulatory distance of output	Semantic distance of output
Primary task	Decide on disjoint searches	Formulate new query from scratch	Perceive the contexts of new searches	Identify results in the display
Ancillary task	Keep in mind the nested searches	Formulate new queries to refine previous results as input	Follow the nested results as part of the context of a single task	Recognize results in the display as part of a nested search

the results of parallel searches separated from each other, the steps required for performing a *search task* do not have an extra impact in the execution and evaluation gulfs.

However, if the search users want to perform is an ancillary task, some design options that would be acceptable for a *primary* search can dramatically increase the execution and the evaluation gulfs. As for the input, semantic distance is larger because users have to keep in mind that the (*new*) *search* is nested in the context of another task. Moreover, users should be able to indicate which part of the results should be refined in a new query, which might increase *articulatory distance of input* if users cannot directly select keywords from the existing results in the display. As for the evaluation gulf, users need to perceive the new results as nested in a previous task which, without proper design, can be challenging and increase *semantic distance of the output*. Finally, *semantic distance of output* can be increased if users are not able to easily recognize the results in the display as being part of a previous task.

## 2.4 Coordination of Multiple Web Searches

Every search task creates a context of use that requires users' attention for formulating a query and interacting with the display. Moreover, according to the users' goals and needs, *Web searches* might become iterative and/or nested tasks. The occurrence of multiple searches running in parallel or at least in short time intervals is an additional source of dispersion of users' attention. However, for determining whether (or not) coordination between multiple search tasks is needed one should consider the users' goals.

Indeed, there are many situations where creating new contexts for a search task is not perceived as a major issue for the users: for example, when users want to test keywords in specific queries, or when users want to compare the results of two searches preserving the original context, or when parallel searches have no dependency with other user tasks. These are typical examples of *primary search tasks*. In these cases, coordination between tasks occurs in the user's mind since keeping the search disjoint corresponds to a user's goal. However, when performing *ancillary search tasks*, what users want is to obtain further details about information already in the display, so there is a clear dependency between what users see on the screen and the results they expect from the new search. It is interesting to notice that every time the user has to perform an *ancillary search task*, this is actually an interruption of a task that cannot be accomplished due to lack of the necessary information. In such cases, creating new contexts might be perceived as misleading as it increases the semantic and articulatory distances of the input (i.e. execution gulf) by dispersing users' attention from the primary task towards an *ancillary search task*. The dispersion of user attention will still increase in terms of output, and requesting users to interact within multiple contexts will break the inner dependency between the users' tasks adding an unexpected interruption. As discussed before [20], resuming a task after an interruption is difficult and may take a long time; interrupted tasks are perceived as harder than uninterrupted ones; interruptions cause more cognitive workload, and they are quite often annoying and frustrating because they distract people from completing their work.

### 3 Inside-in Approach for Ancillary Search Tasks

Hereafter we present the inside-in approach whose ultimate goal is to mitigate the effect of some subtasks that have been identified as difficult to accomplish when performing ancillary searches, such as: to formulate the query, use a new window to interact with the results, and keep a straightforward context for nested search tasks.

#### 3.1 Working Scenario

In order to ground the scenarios around the same application domain, we have chosen to illustrate them with data about co-authorships, as follows:

*“John is expert member of the jury that assesses the research of a Graduate Program in Computer Science. He has to use a Web form which contains the list of ~400 researchers for which he has to provide an assessment based on the number of co-authors and relevant publications. The number of publications and co-authors is required to calculate two important metrics: the researcher’s productivity (accordingly to a formula that takes into account the number of co-authors to estimate the individual effort for the publication) and the size of his networking (as successful scientific collaborations ultimately lead to joint publications). John starts by making a Google search on the Web using the name of the first researcher in the list. Finding the right researcher’s Web page was not easy as the Google search engine returns many entry points including homonymous and some trash pages. After fixing typos and refining the terms of the query, John finds the researcher’s Web page where he can count the number of his publications. For assessing the size of the research network, things are more complicate. John considers two options: i) to create manually a side-list with the names of co-authors; or ii) to look for them in the DBLP web site. John chooses the second option; he types the name of the researcher on the search box of the browser, goes to DBLP web site, scrolls down to reach the zone where co-authors are displayed, and open up the list of co-authors. Now, John is ready to fill in the form, but then he realizes that the DBLP content now occupies the window that previously contained the Web form... For the next 399 researchers John decided to create new tabs for keeping the DBLP search apart from the Web form. Then, he finds out himself being performing repetitive copy-and-paste between tabs, which definitely does not improve his overall performance...”*

#### 3.2 Rationale for the Inside-in Approach Based on the Scenario

In our working scenario, searching co-authors is an ancillary search that complements the user’s main task, which is filling in the Web form. From this scenario, we find some issues that make the following users’ tasks difficult:

- Formulating queries is error-prone (might contain typos) and also time consuming (typing takes time).
- Keywords might be ambiguous, and generic search engines will return broad results. Users may have to scan the list of results until finding the ones that correspond to their goals.

- There are many alternative locations for showing results (including new windows/tabs); choosing the best location for displaying results depends on where the results are meant to be used.
- Some queries might be repetitive; so, saving a few seconds in the time required to complete a single task might represent a huge improvement in the overall task performance at the end of the day.

We claim that the issues raised above can be solved (or at least minimized) with our inside-in approach including the following mechanisms aimed at supporting ancillary search tasks:

- Launching queries from words available in the current Web page can reduce typos. Keywords can be selected with mouse clicks, which is sensibly faster than typing in using a keyboard.
- Ambiguous results are often the result of a broad search. This problem can be reduced by providing specialized queries that operate on specific application domains using user-selected keywords.
- Query results can be shown inside the current page, inline to the selected keywords. This is one of the keystones of the inside-in approach, but queries should be launched on user's demand. If the system systematically launches queries without user's request, the Web page will become polluted by ancillary results, and the benefits of the inside-in approach will be lost.
- Results obtained from ancillary searches should support some kind of interaction to allow users to perform nested queries. This element is important for repetitive tasks, which are often associated to contexts where ancillary searches are required.

The selection of keywords in the text and the use of predefined queries aim at reducing the gulf of execution. This reduction is achieved by minimizing the users' effort in informing keywords to the system (articulatory distance of inputs) and by favoring recognition of keywords and queries rather than imposing the formulation of complete queries (semantic distance of inputs). Predefined queries also help to reduce the evaluation gulf as the results are focused on a particular application domain (semantic distance of outputs). By showing results in the same page and allowing the user to perform nested queries, the inside-in approach helps to reduce the articulatory distance of outputs.

For the sake of simplicity, the scenario used in this section is minimalist and only covers a single level of nested search. However, based on the same principle it would be possible to extend the number of nested search tasks.

## 4 Tool Support for the Inside-in Approach

In this section we present the set of the tools, featuring a framework, which we have developed to demonstrate the feasibility of the inside-in approach. Later on, in this section, we provide some preliminary results we have obtained in an empirical study using our tools.

## 4.1 Architecture

The overall architecture of the framework is briefly illustrated in Fig. 6. It was built upon the concept of Web augmentation [4], which defines strategies for implementing tools that can extend the set of elementary tasks users can do while navigating on the Web. Whilst the full details about the implementation of that framework are out of the scope of this paper, it is interesting to notice that it includes a client-side module and a broker at the server-side.

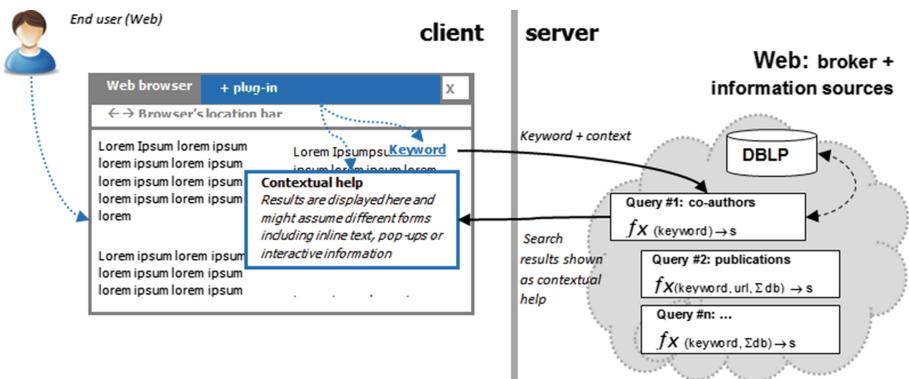


Fig. 6. Overview of the framework architecture for supporting ancillary searches over the Web.

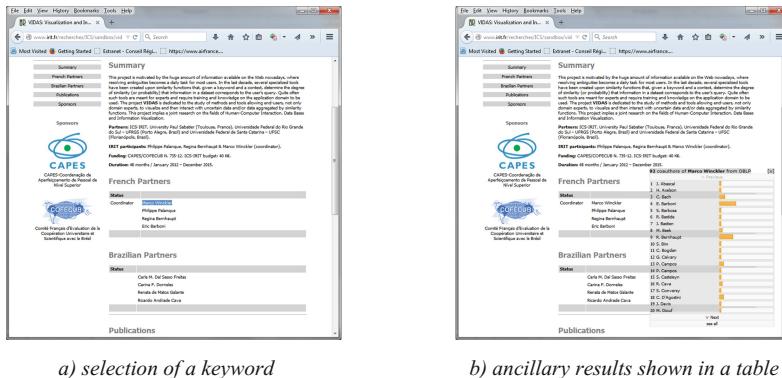
At the server side we have developed a broker, which can be connected to many information sources. This broker was only implemented to illustrate the inside-in approach, and it could eventually be replaced (or connected) with any search engine. The broker contains a set of preprogrammed query functions that are made available to the end users [3]. The set of predefined queries is large, and it aimed at to fit different users' goals for ancillary tasks, for example: finding co-authors, finding publications of a particular author, etc. The number of queries accessible from the client-side can be configured dynamically, but for the purposes of this paper we are just using a specific one, which returns the co-authors of a given researcher. These queries as well as the choice of DBLP database as information source of results are totally arbitrary but justified by the fact that they are related to our case study. Indeed, many different queries are possible to match an ancillary search with diverse users' goals. We suggest that the broker would embed some kind of intelligent behavior for suggesting ancillary searches according to the users' previous search. Nonetheless, in the current implementation of the Web broker we have not integrated any recommender system yet. So, the selection of the predefined search tasks is currently done on the client-side tools.

For the interaction at the client site we have developed a client side module that can be installed as a plugin of the Web browser. This module allows users to select keywords in the current web page and to trigger the queries for ancillary search available at the server side. Once the broker replies, this module modifies the current Web page to display the search results as a kind of contextual help. For that, the DOM

structure of the current Web page is adapted using a set of JavaScript functions, called augmenters [4]. As demonstrated in a previous work [23], adaptations created by augmenters are volatile and do not affect the application running in the Web server. The client-side module can display ancillary data using different interaction techniques, including simple interaction tables displaying a simple list of results or more complex information visualization techniques such as IRIS, which is presented below.

## 4.2 User Interaction with Client-Side Tools

Once the client-side module is running, it grants access to predefined queries available in the broker. Ancillary search queries can be executed by simple selection of words available in any Web pages as shown at Fig. 7.a. With a right-click, the user can select a list of predefined queries and trigger the ancillary search at server side. Query results are returned to the client and displayed in the form of contextual help that superimposes existing contents on the web page, as shown by Fig. 7.b. If no co-author is found for the selected word, the JavaScript function popups an error message informing the user.



a) selection of a keyword

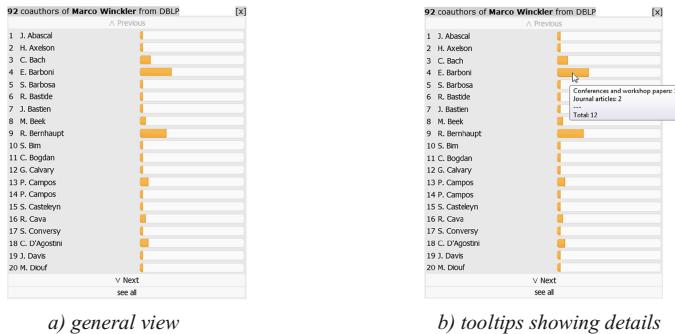
b) ancillary results shown in a table

**Fig. 7.** Overview of user interaction with the client-side tool: (a) the selection of a keyword over the Web page; (b) the ancillary query results shown in a tabular form next to the keyword.

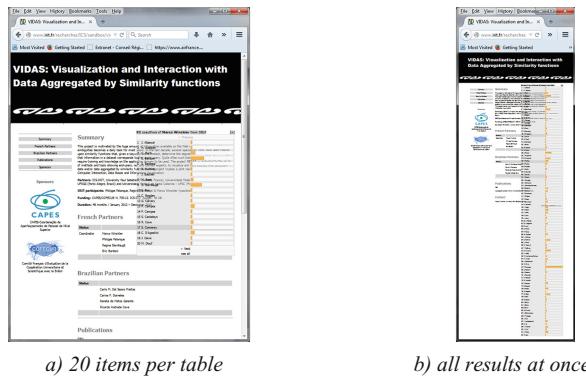
As shown in Fig. 8, the name of the author used as keyword is placed at the topmost row of the table, which also includes the total number of results found. The first column represents co-authors ordered by name, and the second one contains a bar graph depicting the ranking of co-authored papers between the author and each co-author. The details for a ranking can be obtained via tooltips, as shown by Fig. 8.b.

We can also see in Fig. 9.a that the table features 20 items. For more results users can navigate the subsets using the links “*^ previous*” and “*v next*”, or get all results via the option “*see all*”. Notice that the option “*20 results*” requires browsing subsets, while the option “*see all*” requires users to scroll down, as shown in figure Fig. 9.b.

The display of the tabular view for showing ancillary results can be easily customized, for example, by modifying the number of columns to show more results. But



**Fig. 8.** Ancillary results in a tabular form: (a) general view; (b) tooltip with details of ranking.

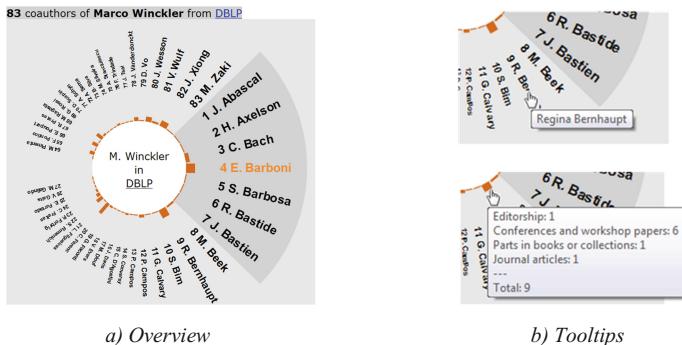


**Fig. 9.** Presentation of tabular view showing: (a) 20 items per table; (b) all results.

more important than the customization of the table view is the possibility of replacing it in the client-side module by another type of visualization that would be a better fit to display ancillary results. Indeed, the inside-in approach we propose does not impose any kind of particular visualization. To illustrate this point, we show in Fig. 10 the same results using a visualization technique called IRIS (which stands for Investigating Relationships between Indexes of Similarity), which features a radial layout and provides different interactors for exploring the set of ancillary query results, including animated transitions and additional tooltips.

### 4.3 Preliminary Results

In order to test our tools we have run a remote empirical study with end-users. The evaluation was designed around a scenario where, given a list of names of researchers in a web page, users should perform ancillary tasks for checking co-authorship at the DBLP. Users were allowed to perform the same tasks by visiting the DBLP web site (<http://www.informatik.uni-trier.de/~ley/db/>) and by using our tools. The list of



**Fig. 10.** IRIS visualization of co-authorship of information from DBLP.

co-authors was obtained by parsing data from DBLP and displaying it using IRIS as shown in Fig. 10, so that users will get exactly the same data.

Users should start at the Web page members of VIDAS' project shown in Fig. 9. From there, users could visit the DBLP web site or use IRIS. Users were offered with three different locations for displaying ancillary search results with IRIS: in a new page/tab, embedded into the Web page layout, shown as a floating window.

The study included an online survey that covers five main chapters as follows: (i) presentation of the study, (ii) occurrence of ancillary search in their daily life; (iii) preference for formulating queries and interacting with results, (iv) preference for the location of the search results in the display, and (v) demographic data.

We have recruited 61 participants via mailing lists and social networks. Most of participants were male (77, 1 %) and, in average 26, 1 years old ( $SD = 5, 7$ ). Among the participants 44.3 % were students, 39, 3 % researchers/professors and 16, 4 % work in the industry. We have got responses from Argentina, Austria, Brazil, France and Spain. They estimated to spend  $\sim 5, 3$  h ( $SD = 4, 1$ ) per week using search engines over the Web, most of which ( $\sim 4, 3$  h per week,  $SD = 3, 9$ ) is spent looking for authors and publications. The amount of time participants perform searches related to authors and publications qualify them as typical users of tools with predefined queries for looking for information, such as searching for co-authors.

The results confirm that typing text to formulate a query is less appreciated than selecting keywords. Most participants found that selecting a term in a web page for launching a query is useful (85, 3 %), and that it improves performance (80, 3 %). Only 18, 1 % of participants prefers typing a text to formulate queries. Considering the alternatives for the location of display, 59 % of participants said to prefer the option embedding results into the current page, while 36, 1 % liked more the design option showing the results in a floating window. Most participants prefer to see ancillary results in the same page (95, 1 %) rather than in a new tab/window (4, 9 %).

Most of the participants clearly pointed out that the options for showing results embedded into the Web document (either the floating window or the option changing the layout) presented the advantage of reducing the interruption created by search engines when showing the results in a new window/tab. The frequency on which such disruption was reported in the comments lets us think that participants really notice the

articulatory distance created when new windows are open. Overall, users did not like the option that shows a new page because of the change of context. This result is compatible with our claims for the inside-in approach. The majority of positive comments were centered on the availability of the additional information right next to the search keyword.

## 5 Related Work and Discussion

Wilson [22] claims that the design of the user interface has an important cognitive impact on tasks performance; thus, search engines should evolve to take into account users' needs. Although these claims are valid, most of the research efforts in the area have been focused on two main subjects: (i) algorithms for improving the accuracy of search engines with respect to many users' concerns and (ii) approaches for improving the visualization of Web pages [19]. For example, Schwarz and Morris [16] describe an information visualization approach for augmenting Web pages with data about the credibility of the ranking proposed by the search engine. While such approach does not help users to formulate better queries, it might help users to better select the pages to visit according to the rank of pages proposed by the search engines. Capra et al. [1] also proposed to augment the user interface of search results by adding images next to the ranking provided by search engines aiming at helping users to make better decisions. These few examples are illustrative of strategies for improving the design and display of the ranking of results from search engines.

In the last decades, several information visualization approaches have been developed for presenting search results coming either from search engines or widely used databases, such as DBLP, ACM DL, IEEE Xplore, etc. Some search engines with built-in visualization tools have also been developed. The first reports presenting and/or discussing visualization of search results date from the late 90's and early 2000's. However, although along the years, many different techniques have been evaluated [17] with results favoring visualizations, the majority of web search engines still provide textual lists ordered by some user or tool specified criteria.

As far as we know, the *inside-in* approach proposed in this paper is an original contribution that can improve users' performance while performing ancillary searches. In some extension, the principles of the inside-in approach can be related to the *focus + context* approach proposed by Card, Mackinlay & Shneiderman [2]. Moreover, it is fully compatible with the Shneiderman's visual information-seeking mantra "*Overview first, zoom and filter, then details-on-demand*" [15].

Although our tools are fully operational, they should be considered a simple proof of concept. Other implementations are possible, indeed, and we suggest that it might require a few extensions of Web browsers to support a native implementation of the inside-in approach. The preliminary results obtained with the survey confirmed our first hypothesis: most users prefer to launch queries directly from the web page by selecting a keyword. This is not a new finding [10] but indicates that we are in the right path. As for the other three hypotheses, they were confirmed: users also prefer search results being displayed through an interactive visualization technique, located near the search keyword. Regarding location, users expressed to prefer the display of results in a way

that does not change their context, this being achieved by two alternatives – displaying the results embedded in the web page, by augmenting it, or displaying them in a floating layer over the same web page.

One of the interesting contributions of the inside-in approach is to allow users to explore the results and perform nested queries that are meant as ancillary-search tasks. The preliminary results of our tools confirm that the semantic and articulatory distances of inputs (*execution gulf*) in the search task are reduced because searching is launched by clicking on a keyword displayed in the Web page. The semantic and articulatory distances of output (*evaluation gulf*) are also reduced when ancillary search results are placed in the same page.

## 6 Conclusions and Future Work

This paper proposed a new perspective for looking at the way search user interfaces can be conceived for helping users to perform ancillary-search tasks on the Web. For that we have proposed the inside-in approach which aims at reducing both execution and evaluation gulfs in the user interaction with search engines. Indeed, one of the key aspects of this approach is to provide a better integration of search results into existing Web pages, where users require complementary information to make their decisions.

Overall the inside-in approach is generic and can be implemented using current search engines such as Google or Yahoo! Nonetheless, it can also be implemented using search engines that are suitable to provide more focused and accurate results about data in a specific application domain. Our framework follows this latter approach as illustrated with the implementation of queries for searching co-authors in the DBLP. While looking up for co-authors might be perceived as a very narrow and specific search, it is noteworthy that it is relevant and frequent in the domains of scientific research, and also is a concern to a large population of researchers, students, teachers, and experts from research funding agencies. Moreover, such specialized characteristic can be tuned and adapted according to specific users' needs. Indeed, the main challenge here remains the identification of relevant queries that are suitable to help users to accomplish their tasks.

Despite the promising results, we know that these are preliminary and there is much work to be done. We would like to measure the distances in the gulfs by performing experiments with direct observation methods. We also intend to proceed with the development of different input and output techniques for performing search tasks since our framework was developed aiming at such studies. Future work also include empirical testing with users in a usability laboratory. This step would allow us to assess user performance when performing the tasks and collect more qualitative data via thinking aloud, which would better explain the user experience factors that influence the use of information visualization techniques for displaying search results. We also plan to develop new techniques for embedding ancillary results into Web pages and then investigate their effect in terms of usability and UX.

**Acknowledgments.** We acknowledge the financial support from the following Brazilian research funding agencies: CAPES/COFECUB, FAPERGS and CNPq. We are also deeply grateful to the anonymous users who willingly served as subjects in our remote study.

## References

1. Capra, R., Arguello, J., Scholer, F.: Augmenting web search surrogates with images. In: Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management (CIKM 2013), pp. 399–408. ACM, New York (2013)
2. Card, S.K., Mackinlay, J.D., Shneiderman, B.: Focus + Context. In: Card, S.K., Mackinlay, J.D., Shneiderman, B. (eds.) Readings in Information Visualization: Using Vision to Think, pp. 307–309. Morgan Kaufmann Publishers, San Francisco (1999)
3. Dorneles, C.F., Gonçalves, R., dos Santos Mello, R.: Approximate data instance matching: a survey. *Knowl. Inf. Syst.* **27**(1), 1–21 (2011)
4. Firmenich, S., Winckler, M., Rossi, G., Gordillo, S.: A framework for concern-sensitive, client-side adaptation. In: Auer, S., Díaz, O., Papadopoulos, G.A. (eds.) ICWE 2011. LNCS, vol. 6757, pp. 198–213. Springer, Heidelberg (2011)
5. Fleming, J.: Web Navigation: Designing the User Experience, p. 264. O'Reilly, Sebastopol (1998)
6. Hassan, A., White, R.W., Dumais, S.T., Wang, Y.-M.: Struggling or exploring?: disambiguating long search sessions. In: Proceedings of PWSMD 2014, pp. 53–62 (2014)
7. Hearst, M.: User interfaces for search, chapter 2. In: Baeza-Yates, R., Ribeiro-Neto, B. (eds.) Modern Information Retrieval: The Concepts and Technology behind Search, 2nd edn. Addison Wesley, New York (2011)
8. Hilbert, M., López, P.: The world's technological capacity to store, communicate, and compute information. *Science* **332**(6025), 60–65 (2011)
9. Marchionini, G.: Interfaces for end-user information seeking. *J. Am. Soc. Inf. Sci.* **43**(2), 156–163 (1999)
10. Marchionini, G.: Exploratory search: from finding to understanding. *Comm. ACM* **49**(4), 41–49 (2006)
11. Martinie, C., Palanque, P., Winckler, M.: Structuring and composition mechanisms to address scalability issues in task models. In: Campos, P., Graham, N., Jorge, J., Nunes, N., Palanque, P., Winckler, M. (eds.) INTERACT 2011, Part III. LNCS, vol. 6948, pp. 589–609. Springer, Heidelberg (2011)
12. Martinie, C.: A Synergistic Models-Based Approach to Develop Usable, Reliable and Operable Interactive Critical Systems. Ph.D. thesis presented on December 25th 2011. Université Paul Sabatier (2011). <http://www.irit.fr/~Marco.Winckler/martinie THESE2011.pdf>
13. Norman, D.: The Psychology Of Everyday Things. Basic Books; 1 edition (June 13, 1988). Basic Books. ISBN 978-0-465-06710-7 (1988)
14. Norman, D.A., Draper, S.W. (eds.): User Centered System Design: New Perspectives on Human-Computer Interaction. Lawrence Erlbaum Associates, Hillsdale (1986)
15. Paternò, F., Mancini, C., Meniconi, S.: ConcurTaskTrees: a diagrammatic notation for specifying task models. In: Proceedings of Interact 1997, pp. 362–369. Chapman & Hall (1997)
16. Schwarz, J., Morris, M.R.: Augmenting web pages and search results to support credibility assessment. CHI 2011, pp. 1245–1254

17. Sebrechts, M., Vasilakis, J., Miller, M., Cugini, J., Laskowski, S.: Visualization of search results: a comparative evaluation of text, 2d and 3d interfaces. Proceedings of the 22nd ACM SIGIR Conference on Research and Development in Information Retrieval, pp 3–10 (1999)
18. Sutcliffe, A.G., Ennis, M.: Towards a cognitive theory of information retrieval. *Interact. Comput.* **10**, 321–351 (1998)
19. Suzuki, E., Ando, S., Hirose, M., Jumi, M.: Intuitive display for search engines toward fast detection of peculiar WWW pages. In: Zhong, N., Liu, J., Yao, Y., Wu, J., Lu, S., Li, K. (eds.) Web Intelligence Meets Brain Informatics. LNCS (LNAI), vol. 4845, pp. 341–352. Springer, Heidelberg (2007)
20. ter Beek, M.H., Faconti, G.P., Massink, M., Palanque, P.A., Winckler, M.: Resilience of interaction techniques to interrupts: a formal model-based approach. In: Gross, T., Gulliksen, J., Kotzé, P., Oestreicher, L., Palanque, P., Prates, R.O., Winckler, M. (eds.) INTERACT 2009. LNCS, vol. 5726, pp. 494–509. Springer, Heidelberg (2009)
21. Yates, R.B., Neto, B.R.: Modern Information Retrieval: The Concepts and Technology behind Search, 2nd edn, p. 944. ACM Press Books / Addison-Wesley Professional, New York (2011)
22. Wilson, M.L. Evaluating the cognitive impact of search user interface design decisions. In: EuroHCIR 2011, pp. 27–30 (2011)
23. Winckler, M., Gaits, V., Vo, D.-B., Firmenich, S., Rossi, G.: An approach and tool support for assisting users to fill-in web forms with personal information. In: Proceedings of the ACM SIGDOC 2011, Pisa, Italy, pp. 195–202. ACM, New York (2011)

# Using Affinity Diagrams to Evaluate Interactive Prototypes

Andrés Lucero<sup>(✉)</sup>

Mads Clausen Institute, University of Southern Denmark, Kolding, Denmark  
lucero@acm.org

**Abstract.** Affinity diagramming is a technique used to externalize, make sense of, and organize large amounts of unstructured, far-ranging, and seemingly dissimilar qualitative data. HCI and interaction design practitioners have adopted and used affinity diagrams for different purposes. This paper discusses our particular use of affinity diagramming in prototype evaluations. We reflect on a decade's experience using affinity diagramming across a number of projects, both in industry and academia. Our affinity diagramming process in interaction design has been tailored and consists of four stages: *creating notes*, *clustering notes*, *walking the wall*, and *documentation*. We draw examples from eight projects to illustrate our particular practices along these four stages, as well as to ground the discussion.

**Keywords:** Interaction design · KJ method · Evaluation · Analysis

## 1 Introduction

Researchers have recently been looking into and studying different (design) methods and reflecting on how they are used in practice by the human-computer interaction (HCI) and interaction design communities. Most of these studies have looked into methods that were originally conceived within and are closely related to design practice, such as probes [22, 32], workbooks [6], and mood boards [24]. First introduced in the 1960s, affinity diagramming (or the KJ method) [2, 13, 16] has its origins in anthropology and social science and has widely been accepted within HCI research. Affinity diagramming is a technique used to externalize, make sense of, and organize large amounts of unstructured, far-ranging, and seemingly dissimilar qualitative data [12]. Common uses of affinity diagramming include analyzing contextual inquiry data [2, 13] clustering user attributes into profiles [21] or requirements [1], problem framing and idea generation [7, 30] and prioritizing issues in usability tests [9].

In this paper, we reflect on a decade's experience using affinity diagramming to evaluate interactive prototypes. Our affinity teams usually consist of two researchers who collect data from 10 to 24 participants (i.e., observations of use during a task, and semi-structured interviews), independently write affinity notes (i.e., 500 to 2500 notes), and jointly analyze the data (i.e., build an affinity diagram) over a period of two to three weeks. To better suit small to medium interaction design projects in industrial and academic contexts, we have tailored and scaled down Beyer and Holtzblatt's six stages of contextual design [2, 13] to four stages. First, when *creating notes*, we embrace the

affordances of paper [18, 28] and build affinity diagrams with physical paper by producing handwritten sticky notes. Second, when *clustering notes*, we invite team members to go through each other's notes in sequence, to avoid ownership issues and to create a better understanding of the context when an observation of use is made. We also avoid using interview questions to structure the data, letting overarching topics naturally emerge [4] from it. Third, in *walking the wall*, we take advantage of color-coded sticky notes (i.e., one color per participant) to check at a glance whether enough people have raised an issue. We also discuss practices related to pruning the wall, which include merging, arranging, and removing note clusters. Finally, in *documentation*, we pick relevant user quotes and count notes to communicate and quantify our main findings. The main contributions of this paper include: a systematic analysis of affinity diagramming use for prototype evaluations in HCI and interaction design over an extended period; an adaptation of earlier affinity diagramming techniques such as the ones described by Beyer and Holtzblatt [2, 13] which have been tailored to suit small to medium projects; and a discussion on practices that are relevant for general affinity diagramming.

This paper is structured as follows. We begin by discussing related work on affinity diagrams, how they are used in HCI and interaction design, and existing support tools. We then take real-life examples from eight industrial research projects to illustrate the four stages of our particular use of affinity diagramming for interactive prototype evaluations. We close by discussing what we have learnt by adapting affinity diagramming to our own practices, followed by conclusions.

## 2 Related Work

### 2.1 Affinity Diagramming and the KJ Method

Affinity diagramming is a technique used to externalize, make sense of, and organize large amounts of unstructured, far-ranging, and seemingly dissimilar qualitative data [12]. Japanese anthropologist Jiro Kawakita devised the KJ method [16] (on which affinity diagramming is based on) as a tool for use in anthropology to synthesize idiosyncratic observations of raw data obtained through fieldwork to find new hypotheses. In Japan, the KJ method has become popular as a systematic approach to problem solving in fields such as research, invention, planning, and education.

Affinity diagrams (or KJ method charts) are a wall-sized paper-based hierarchical representation of data. The KJ method consists of four basic steps [16]. In *label making*, the main facts or issues in relation to the data are captured onto separate pieces of paper (or sticky notes). Rather than grouping notes in predefined categories, affinity diagrams are built from the bottom up. Therefore, in *label grouping*, individual notes are shuffled and spread out on a table, and then read several times. After interpreting and considering their underlying significance [9], individual notes are put up on a large empty table or blank wall one at a time, forming teams of labels (or clusters) that are iteratively rearranged [29]. ‘Lone wolves’, or notes that do not seem to fit in the existing clusters, are left aside for later use. Clusters are then given titles (or named) and, if working on a table with pieces of paper, all notes are put together in a pile with the title

clipped on top. Cluster names and ‘lone wolves’ are read and grouped into more abstract groups giving rise to general and overarching themes [20]. In *chart making*, the resulting clusters are spatially arranged and transferred to a large sheet of paper, where they are annotated using symbols and signs (i.e., connection, cause and effect, interdependence, contradiction) to show the relationships between groups. Finally, in *explanation*, the resulting chart is first verbally explained and then described in writing.

Although conceived for individual use, the KJ method is well suited for collaborative data analysis, supporting parallel work and creation of a shared interpretation of the data. The greater-than-human-sized space often used allows people to simultaneously view, discuss, and modify the artifact [18]. The simple skills needed to fill a table with pieces of paper, or a wall with sticky notes, and move notes around to suggest new associations make affinity diagramming (or the KJ method) a tangible, easy, and approachable way to look into complex data.

## 2.2 Different Uses of Affinity Diagrams

**Contextual Inquiry.** Beyer and Holtzblatt [2, 13] adapted the original KJ method to analyze observational and interview data gathered during contextual inquiries [5, 9, 20, 21]. Affinity diagramming is often used as a starting point for design [20], helping keep design teams grounded in data as they design [9]. The contextual design process begins with *contextual interviews*, where field data from between four and six different work sites is gathered in an attempt to understand work practice across all customers. Second, *interpretation sessions and work modeling* allow every team member to experience the interviews and capture key points (affinity notes), after which five models of the user’s work are created (i.e., flow, cultural, sequence, physical and artifact models). Third, *consolidation and affinity diagram building* consist of merging the data from the five models and the affinity notes to represent the work of the targeted population. Fourth, in *visioning* the team reviews the consolidated data by walking the data and then running a visioning session of how the user’s work will be streamlined and transformed. Fifth, *storyboarding* consists of fleshing out the vision in detail using hand drawn pictures and text. Finally, *paper prototypes and mockups interviews* consist of designing user interfaces on paper and testing them with users.

While Beyer and Holtzblatt’s use of affinity diagrams is aimed at the early stages of the design process, our work looks at how affinity diagrams can provide support to analyze interactive prototype evaluations in the later stages of the design process. Moreover, stages such as *contextual interviews*, *work modeling*, *storyboarding*, or *paper prototypes and mockups interviews* are no longer relevant once interactive prototypes are in place and ready to be evaluated, as those activities should have happened earlier in the process.

**User Profiles and Requirements.** At the start of developing a new software or product, user profiles (or personas) are constructed to provide a direction to the whole process [21]. Team members create a list of audience attributes on sticky notes, thus bringing their own views on the user. These attributes are then collectively clustered into three to eight user profiles. Notes that do not fit in existing clusters create new

ones. Clusters are discussed and notes are moved around until everyone agrees on them. Sometimes it is the customers themselves (i.e., experts in a given field) who create an affinity diagram to specify requirements for a new system [1, 4]. Starting from the design brief, experts write down fairly succinct statements (i.e., requirements, needs, wishes and hopes) on sticky notes capturing the features, properties and expected behaviors of parts of the new system. Requirements are clustered into common themes, while near-duplicate ones are discarded.

**Problem Framing and Idea Generation.** In design practice, affinity diagrams are used to analyze a design problem or to create first design solutions [4, 7]. With a given design problem in mind, designers write down words, short sentences, or create small sketches on sticky notes to stimulate diversity in ideation phases. Ideas are then clustered to identify common issues and potential solutions, ultimately helping to frame the design problem. Used to achieve consensus and understanding among participants in a discussion [30], a radial affinity diagram is usually built on a table. A key problem or theme is first placed in the center of the base sheet. Participants then take turns in placing their cards on the table, aligning similar cards in a spoke-like fashion. Visual connections between similar cards can be created using lines of paper clips. After discussion, participants fix the card positions and links. In web design, affinity diagramming is used as a form of collaborative sketching [18] to create sitemaps with the structure of a website. Information architects and visual designers first collect ideas about what should be in a website onto sticky notes and then arrange them on the wall into categories, usually on a whiteboard. Sitemaps can grow fast and end up including 200 to 300 notes. Visual designers sketch page designs directly on empty spaces of the whiteboard.

**Usability Tests.** In usability testing, affinity diagrams are also used to help teams prioritize which issues will be fixed and retested [4, 9]. At the start of a usability test session, the team assigns a sticky note color to each participant and watches as they perform tasks from an observation room. Observations and quotes are captured on the notes, which are put up on a wall. Common interface issues and problems will emerge. Several colored notes in one issue will indicate that many people experienced a similar problem and should probably be addressed first.

In HCI and interaction design, affinity diagrams have also been used to analyze (post-task) interview data from interactive prototype studies [3, 31]. In this paper, the novel and particular use of affinity diagrams in prototype evaluations to analyze both observations of use while participants perform a task and (post-task) interview data is discussed [17, 23, 25, 26].

### 2.3 Affinity Diagramming Support Tools

Software tools have been available to create affinity diagrams. CDTools<sup>1</sup> was a software package to support the Contextual Design process offered by InContext Design. PathMaker<sup>2</sup> is a split screen interface that allows recording, dragging, and grouping

---

<sup>1</sup> CDTools. <http://incontextdesign.com/process/cdtools/>.

<sup>2</sup> PathMaker. [www.skymark.com/pathmaker/tour/brain.asp](http://www.skymark.com/pathmaker/tour/brain.asp).

ideas into affinity sets. StickySorter<sup>3</sup> allows working visually with large collections of notes. Koh and Su [19] created an affinity diagram authoring environment that provides an infinitely large workspace to post and organize notes, dynamic group layout and repositioning, meaningful zoom levels for navigation (i.e., to fit note, group of notes or selection area), ways to save, restore and export diagrams (JPEG), as well as ways to search for notes and groups (i.e., by text matching and related words). The main disadvantages of these tools include a lack of support for collaboration (i.e., single user) and having to do things in a certain structured way [10].

Support for affinity diagramming has also been available in prototype form. The Designers Outpost [18] combines the affordances of paper and large physical work-spaces with the advantages of digital media. People write on sticky notes with a normal pen, which are then added to an electronic whiteboard used as a canvas. Using computer vision, the system tracks the position of the notes, which can be physically moved around or removed from the board. Physical notes can be structured, by drawing lines from one note to another, and annotated with digital pens. AffinityTable [8] replicates the benefits of physical affinity diagramming (i.e., copying, clustering, piling and collecting) and enhances current practices (i.e., highlighting, focusing, searching, retrieving images) by combining one vertical and one horizontal display, plus digital pen and paper. Multi-touch gestures and physical tokens provide input to the interactive table. The GKJ system [30] provides a way to digitize affinity diagrams and the process of building them. Using wireless Anoto-based pens, the system records annotations on cards and on a base sheet. Pen gestures are used to determine the position and orientation of a card, as well as to group/ungroup clusters of cards. Time stamped gestures allow visiting the history of the affinity diagram process, and also serves as an undo function. A PC editor allows further editing the virtual diagram. Harboe et al. [11] proposed a distributed system of digital devices (i.e., mobile phones, tablets, stationary camera, and projector) to augment affinity diagramming while aiming to support existing paper-based practices. Using a Magic Lens metaphor, paper notes tagged with unique QR codes can be tracked with the camera of a mobile phone or tablet. Once recognized, the note can be augmented with additional metadata information, which is projected on top of the physical note. The prototypes discussed here aim to augment affinity diagramming, however Klemmer et al. [18] report that some digital features have the potential to interrupt the designers' creative flow and can be considered distracting (i.e., too many things flashing).

Despite the pervasiveness of new technologies, paper remains a critical feature of work and collaboration [27, 28]. Luff et al. [28] discuss some of the affordances of paper that seem critical to human conduct. Paper is *mobile* as it can easily be relocated and juxtaposed with other artifacts, and *micro-mobile* as it can be positioned in delicate ways to support mutual access and collaboration. Paper can be *annotated* in ad hoc ways, allowing people to track the development of the annotations and recognize who has done what. Paper is *persistent* [18], retaining its form and the character of the artwork produced on its surface. In addition, paper allows people to simultaneously see its contents from different reading angles, and it can become the focus of gestures and

---

<sup>3</sup> StickySorter. [www.youtube.com/watch?v=Wg0WYcYlls0](http://www.youtube.com/watch?v=Wg0WYcYlls0).

remarks [27]. The affordances of paper have played a key role in our practices with affinity diagrams, including our preference to use physical paper to digital alternatives, as well as to manually write notes on sticky notes.

### 3 Eight Affinity Diagrams

We reflect on a decade's experience using affinity diagramming to evaluate interactive prototypes, both in industry and academia. Real-life examples from eight industrial research projects where the technique has been used (see Table 1 for an overview) will help illustrate how affinity diagrams are used in practice, as well as to ground the discussion.

**Table 1.** Overview of eight affinity diagramming cases.

Prototype	Evaluation type	Evaluation participants	Note takers	Note technique	Affinity notes	Affinity team size	Notes on wall	Notes discarded
A <i>MindMap</i>	Group (2 × 3)	6	2	Mix	258	2	232	10 %
B <i>Pass-TheM-Around</i>	Group (5 × 4)	20	2	Mix	811	3	740	9 %
C <i>MobiComics</i>	Group (3 × 9)	27	2	Printed	2433	5	2243	8 %
D <i>EasyGroups</i>	Group (6 × 4)	24	2	Mix	1096	2	992	9 %
E <i>FlexiGroups</i>	Group (4 × 6)	24	2	Mix	715	2	673	6 %
F <i>Image Space</i>	Individual	10	2	Handwritten	505	2	391	23 %
G <i>Twisting Touch</i>	Individual	24	2	Handwritten	1037	2	630	39 %
H <i>NotifEye</i>	Individual	13	2	Handwritten	1276	2	946	26 %

Prototypes A to E [25] were related to groupware, exploring the use of mobile phones for collaborative interactions in different physical and social use contexts (i.e., office work, media consumption at home, public expression in a pub, and general group formation). These prototype evaluations were conducted with different numbers of participants (i.e., between six and 27 people) in groups of varying sizes (i.e., between three and nine people per group). Prototypes F, G, and H, on the other hand, were evaluated individually (between 10 and 24 participants per evaluation). Prototype F [23] was a social network service built around sharing personal photos. Prototype G [17] allowed using a flexible handheld interface to provide input for interaction. Finally, Prototype H [26] explored the use of interactive glasses to provide notifications on the go. Most of these evaluations were conducted in controlled lab environments, except for prototypes C, F, and H, which took place in public spaces.

In all prototype evaluations, two researchers independently made notes as they watched videos of participants performing an interaction task and a semi-structured interview. Handwritten notes were created for all prototypes but one, i.e. C, which used digital notes printed on label templates only. In addition, digital notes printed on sticky notes or on paper (plus removable tape) were produced for four prototypes (i.e., A, B, D, and E), resulting in a mix of handwritten and digitally created notes. The size of the team that built the affinity diagram ranged from two to five people, and always included

the same two researchers who made the notes in the first place. The resulting affinity walls differed in size, ranging from 232 to 2243 notes. Around 15 % of the notes were discarded, except for two cases, prototypes G (39 %) and H (26 %).

## 4 Affinity Diagramming Process

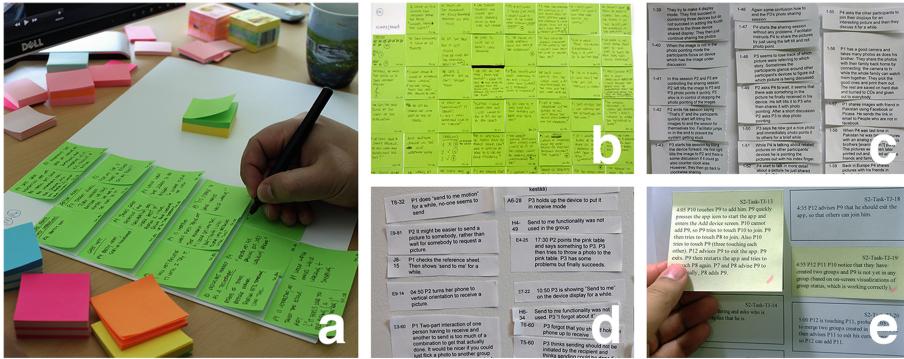
We outline stages and properties of our particular use of affinity diagramming for interactive prototype evaluations. As was mentioned earlier, Beyer and Holtzblatt's [2, 13] use of affinity diagrams is intended for the early stages of the design process, and some of their stages are not relevant for prototype evaluations. We have tailored their process to better suit small to medium interaction design projects in industrial and academic contexts. More specifically, we have combined their first two stages (i.e., *contextual interviews* and *interpretation sessions and work modeling*) into *creating notes* by placing prototype evaluations at the core, and removing the work modeling activity. Their third and fourth stages (i.e., *consolidation* and *affinity diagram building* and *visioning*) are closely related to *clustering notes* and *walking the wall*, respectively. Finally, their last two stages (i.e., *storyboarding* and *paper prototypes and mockup interviews*) have been replaced by *documentation*. Our process consists of four stages: *creating notes*, *clustering notes*, *walking the wall*, and *documentation*.

### 4.1 Creating Affinity Notes

This first stage of the process starts with the actual evaluation of the prototype. After carefully planning the main research questions, internal ethics committees and privacy reviews, consent forms, the introduction, the task, coffee breaks, the semi-structured interview questions, the debriefing, and rewards, we collect data over a period of one or two weeks (depending on the number of participants). Each evaluation session usually lasts between one and two hours. As a result, we end up with 6 to 24 h of video to analyze, both from observations of use during a task, and from the semi-structured interviews. Interpretation sessions are then conducted within 48 h after the prototype evaluations. Two researchers with mixed background (i.e., a designer and a psychologist, or a designer and a computer scientist) independently make notes as they watch videos of an interaction task and a semi-structured interview. Affinity notes typically include handwritten text, but can also comprise drawings and annotations (Fig. 5). The number of affinity notes can vary between 500 and 2500, depending on the number of interviews, their duration and the level of detail captured. It usually takes us twice as much time to write affinity notes as the length of the videos.

**Handwritten Sticky Notes.** We begin by placing a stack of sheets of A3 paper on our desk (Fig. 1a). These sheets are used as a canvas onto which to stick notes. A3 is a comfortable format to work with while writing affinity notes; it is large enough to put several sticky notes on it, and small enough to fit on a desk in front of a computer. Traditionally, standard  $3 \times 3$  or  $5 \times 3$  inch ( $7.6 \times 12.7$  cm) Post-it notes have been used to write affinity notes. However, we have found those sizes to unnecessarily increase the overall size of the affinity wall and thus the amount of walking for the affinity team

members. Moreover, larger affinity notes invite more verbose expressions from the note takers. Therefore, we use the smaller  $2 \times 2$  or  $2 \times 1.5$  inch ( $5.1 \times 3.8$  cm) sticky notes, which are available in different brands, colors, and slightly different sizes (Fig. 1a). These smaller sticky notes fit in a grid of 8 by 6 notes, totaling 48 notes per sheet.



**Fig. 1.** Creating affinity notes. (a) Writing notes on sheets of A3 paper, (b) handwritten sticky notes, (c) digital notes printed on paper, cut with scissors and attached with tape, (d) digital notes printed on labels, and (e) manually printed sticky notes.

We assign a sticky note color for each participant (Fig. 1b). In case of group evaluations (e.g., prototypes A to E on Table 1), we define one color for each group of participants. With smaller groups of people (e.g., three participants), we try to treat them as individuals as much as possible and thus assign three sticky note colors, one for each. Using different note colors for each participant allows us to tell how many people raise a certain issue by glancing at each category.

Blank sticky notes are arranged (usually in columns) and are given a unique identifier consisting of a participant or session number, followed by a running sequence number (e.g., P1\_01 or S1\_01), which is handwritten in the lower-right corner. Note takers do not need to identify themselves on each note as their handwriting provides a quick way to know the author of a given note. We draw a single thick black line to separate task from interview data (Fig. 1b). We place two or three blank sticky notes on the A3 canvas before starting the video to take notes. To optimize our use of time, subsequent notes are then added and identified in parallel during natural transitions or silent moments in the video. However, some parts of the video do require us to explicitly pause and rewind the video to capture a note in more detail. Those breaks also provide opportunities to add new blank notes to the canvas. Once a sheet of A3 paper is filled with 48 affinity notes, that sheet is moved to the back of the stack. Working with a stack of sheets allows the note taker to quickly consult whether something has been missed or captured earlier. Once all videos have been analyzed, the sheets of A3 paper filled with affinity notes can now be easily transported to the affinity room.

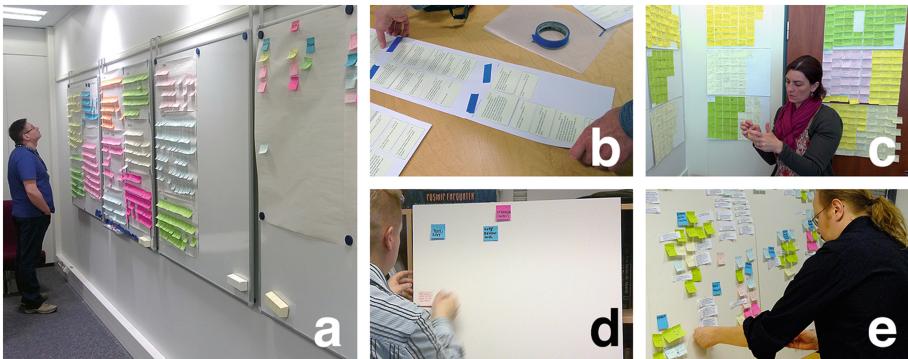
**Digital Notes Printed on Paper, Labels or Sticky Notes.** We have also experimented with using different digital note types in our evaluations of prototypes A to E (Table 1).

First, digital affinity notes are typed on personal computers, then printed on paper, and each note is individually cut with scissors [10] (Fig. 1c). In an attempt to provide a similar function and feel as sticky notes, removable tape, blue painter's tape, or yellow masking tape are used so the notes can be easily attached and moved around the affinity wall. The extensive manual work needed to produce such notes can easily delay the start of building the affinity wall by a couple of hours. Second, label templates (e.g., Avery®) are another alternative to printing digital affinity notes (Fig. 1d). Since the entire surface of the note then becomes adhesive, it produces mixed results as we have had difficulties removing the notes from the wall depending on the wall finishing materials (e.g., paint, wallpaper, wood, glass, etc.). Third, affinity notes can be laser printed on sheets of Post-it® notes. However, priced at almost US\$1 per note and available in one-size ( $3 \times 3$  inches), this alternative is both expensive and impractical. We have come up with a cheaper way to print digital affinity notes on standard sticky notes (Fig. 1e). *MixedNotes* [14] is a software tool that imports text from any text editor, recognizes blank lines as separates notes, adds a unique sequential identifier, optimizes the size of the note to be printed based on the amount of text, and shows an on-screen preview of how notes will be printed on paper. *MixedNotes* first prints a background sheet, a template onto which sticky notes manually cut in three different sizes are attached, and then the same sheet must be put back into the printer for the digital notes to be printed on the sticky notes. Despite claims that digital notes could improve current practices (i.e., faster note-taking, searching), we have not found this to be the case (e.g., search option only used once).

## 4.2 Clustering Notes

**Selecting and Preparing a Room.** Holtzblatt et al. [13] stress the importance of getting a dedicated team room for the duration of the project to avoid wasting time finding another room, packing up materials, and relocating half way through the process. However, not all organizations have a room that can be blocked for one or two weeks. We have used different strategies to secure such a dedicated space for a couple of weeks both in companies and universities. For example, we have reserved an internal meeting room for two weeks (Fig. 2c), used our own shared office (Fig. 2d), used a common innovation space outside our premises, temporarily repurposed a usability lab (Fig. 2a), and used an office space that was emptied before major renovations (Fig. 2e). Room sizes have ranged from a  $2.3 \times 3.6$  m meeting room (Fig. 2c) to a  $6 \times 8.4$  m usability lab (Fig. 2a). The room should provide plenty of wall space for the affinity to spread out [13].

We begin by mounting the sheets of A3 paper containing the affinity notes on the wall (between 1.2-2 m from the floor) using masking tape (Fig. 2b). We then line up the remaining wall space with white flipchart sheets, butcher paper, or statically charged polypropylene film onto which the note clusters will form. In addition, we sometimes put white cardboard panels on tables (Fig. 2d). When using smaller rooms, we have even used the windows and the door (Fig. 2c). Our ideal room (Fig. 2a) has moveable whiteboards that we line up with white flip chart sheets of paper held up by magnets. Such a space allows us to rearrange parts of the wall, be it entire whiteboard panels or



**Fig. 2.** Clustering notes. (a) A typical room at the start of clustering with notes on the left and right, and empty space in the middle, (b) preparing the room by putting up sheets of affinity notes on the wall, (c) going through notes individually, (d) forming first clusters on a white cardboard panel, and (e) clusters have initial names and a few notes below them.

just a couple of sheets. In such a space, we mount affinity notes towards the left and right wall edges, leaving an empty space in the middle for the affinity wall.

**Affinity Diagramming for Prototype Evaluations.** When building an affinity wall to analyze prototype evaluation data, the two note takers who created the affinity notes are already familiar with the data since they have seen the videos for all sessions. In cases where the affinity team has included more than two persons (e.g., prototypes B and C), although these people have not created affinity notes, they have been present earlier as observers during the actual prototype evaluations. Therefore, unlike Holtzblatt et al. [13], at the start of building the affinity we do not break the notes up into piles of 20 (i.e., to make it less intimidating). Instead, we invite people to start by going through each other's notes sequentially to avoid potential ownership issues. Our data also more heavily depends on the details of the context when an observation of use was made, therefore we do not mix up the notes (i.e., mix up users). For example, when going through the first part of the evaluation (i.e., the task), it is important for us to be able to identify whether, e.g., a certain function was successfully triggered or a group was having problems on the first, second or third try. An isolated note saying, “*they are having difficulties completing the task,*” will be interpreted differently whether this is happening at the start or by the end of the task.

Despite the seemingly structured way of analyzing the interaction observations, when going through the second part of the data (i.e., the semi-structured interview questions) we avoid as much as possible using the interview questions to structure the data. Instead, we let overarching topics naturally emerge from the data.

**Building the Affinity.** At the start of building the affinity, team members start by reading each other's notes in silence (Fig. 2c). People will pick notes that raise important issues in relation to the prototype and begin forming rough clusters (Fig. 2d). Once a couple of clusters have been created with a few notes, people will begin verbally coordinating where certain notes are being clustered. Questions such as, “*where are you putting the notes related to this issue?*” will begin to emerge. Thus we

alternate between moments of silence and moments of discussion, the former near the start of the process, the latter as the affinity progresses. Clusters with a few notes below them are initially named and labeled with a blue note (Fig. 2e). These clusters are in turn grouped into more abstract groups labeled with pink notes.

### 4.3 Walking the Wall

**Discussing and Pruning the Wall.** After the team has completed a first round reading all notes, roughly between a third and half of the affinity notes will have been moved from the A3 sheets of paper to the affinity wall, thus forming note clusters (Fig. 3a). Early rounds of discussing the wall will then concentrate on communicating the emerging clusters with the team, checking if these clusters fail to cover some important general topics, and identifying overlapping clusters that could potentially be merged. As a result of these discussions, the teams will agree on an initial set of clusters and (blue and pink) labels. Drawing arcs [12] and sticking stripes of tape [15] between clusters can be used to visually show related parts of the affinity diagram. A note that belongs in two clusters can be duplicated, or split by ripping the paper and adding tape.

The team may also decide at this point to define a tentative cluster hierarchy. On one hand, fixing a hierarchy this early on in the process tends to limit the bottom-up nature of the process. On the other hand, shifting panels around late into the process to define a hierarchy can have a detrimental effect. Social and spatial awareness of the affinity diagram is an important part of building a cognitive model of the data [4, 10, 16]. Moving parts of the affinity wall can create confusion and lead to a waste of time when people are trying to find where to place a given note.



**Fig. 3.** Walking the wall. (a) Affinity notes have been moved the A3 sheets of paper to form the affinity wall, (b) people slowly start discussing the contents of each category, (c) moving several notes at a time when there are enough categories, (d) verbalizing the act of moving a note to the wall, and (e) user statements in first person are written on larger yellow sticky notes (Color figure online).

In later rounds of discussing the wall, the team will more closely inspect the contents of each cluster (Fig. 3b). Specific notes that are unrelated to a cluster may be put on a different cluster, set aside to potentially create a new cluster, or even back in its original location on the A3 sheets of affinity notes. Similarly, clusters may be merged, moved to a different location, or can altogether disappear. Pruning the wall thus includes merging clusters, arranging the cluster hierarchy, and removing notes (and clusters) from the affinity wall.

**Adding Notes to Existing Clusters.** As was mentioned earlier, the affinity team will switch back to reading affinity notes from the A3 sheets of papers in between rounds of discussing and pruning the wall. As the affinity wall progresses, team members will be more familiar with the existing clusters and have a better understanding of where a given note could be put up on the wall. As a result, people will evolve from moving one note at a time to the wall, to several simultaneously (Fig. 3c). Each finger can represent a cluster, and several notes can be placed on each finger, and thus ten notes can be moved to the wall at the same time. Another phenomenon that we have observed at these later stages of the process is that team members are more open to verbalize the act of moving a given note up to the wall (Fig. 3d). – “*Listen to this note [reads the note aloud]. It goes here.*” – “*Yes.*” Such actions are performed to confirm that existing categories remain valid also after new bits of data are added to the wall.

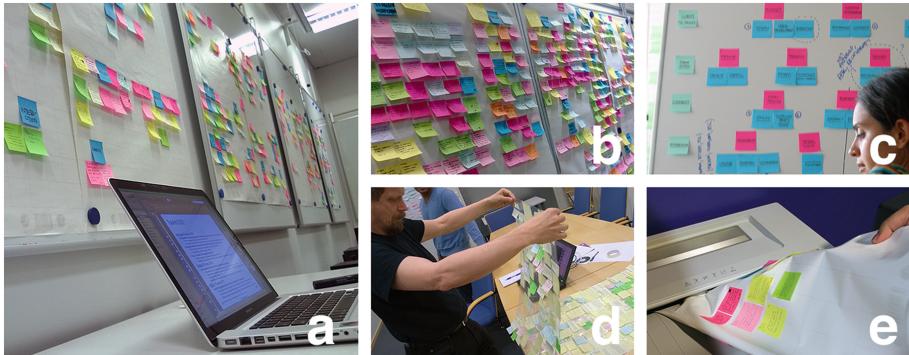
Each round of discussing and pruning the wall, and adding notes to existing clusters often takes between 60 and 160 min. Besides alternating between these two different types of activities, it is important to have 30 to 60 min breaks in between rounds. Affinity diagramming can be a mentally demanding activity [13], especially when it is carried out for three to five days in a row. Natural breaks such as lunchtime or taking a coffee, but also an unrelated meeting can provide a much-needed time for the mind to rest and think about something else, other than the ongoing analysis.

**Finalizing the Wall.** Once there are no more useful notes left on the wall, typically the last 15 % of the notes (Table 1), we read each category note by note and make sure each note belongs to that category. We also make sure there are enough notes and more than two people raising an issue in each category. We then check that every blue and pink note still makes sense within the hierarchy. For each blue label, a succinct user statement describing the issue that holds all individual notes together is written on a large sticky note [2] (Fig. 3e).

#### 4.4 Documentation

**Creating a Digital Record of the Affinity Wall.** Keeping a digital record of the finalized affinity wall (Fig. 4b) allows sharing the results across sites. Miura et al. [30] indicate that in the typical KJ method, there is one participant who digitizes the outcome of the diagram by inputting all card content and arranging the card structure using a mouse and keyboard. Similarly, we assign a scribe [18], a member of the affinity team who saves the resulting info. Using PowerPoint or Word, a document with the hierarchical category structure and its descriptions is created (Fig. 4a). Depending on the number of affinity notes and detail of the documentation, roughly one to three

PowerPoint slides are made for every pink label. Each slide contains the name of the pink and blue labels, the user statement written in first person that describes all notes under a blue label, plus a selection of the most representative user quotes on that particular issue (typically between 1 and 4 user quotes per blue label). Picking relevant user quotes, ones that capture the essence of what participants tried to tell us, will play an important role in communicating the main (positive and negative) findings of the study to different stakeholders, improving existing designs, and further disseminating the end results (e.g., publication at an academic conference).



**Fig. 4.** Documenting the wall. (a) Digitally recording the wall using a laptop, (b) a final affinity wall with pink and blue labels, (c) building the final category structure on a whiteboard, (d) cleaning the room to its original state, and (e) shredding the affinity wall (Color figure online).

In addition to making a digital version of the affinity diagram, we take (high-resolution) digital photographs of each panel. Each photo usually contains one pink note, two to five blue notes, plus their corresponding affinity notes that make up those categories. We sometimes also take close-up panel shots (e.g., if we need to leave the room and save the results to a digital document at a later time). Proper lighting conditions and (color) contrast between the text and note should be considered when photographing panels [19]. To visually assist the wall documentation process, we sometimes build the final hierarchical structure of categories on a whiteboard (Fig. 4c).

**Quantifying Observations of Use and Raised Issues.** Another particular use of affinity walls for interaction design that we have developed over the years is to count the total number of notes and the number of people that raised an issue. Counting the total number of notes allows us to check how frequently the participants mentioned an issue or topic. When creating the final affinity wall hierarchy, the overall note numbers for each pink and blue label provide us with an additional way to prioritize one topic over another. Special care should be taken to identify if a category with a large number of notes consists of one or two people mentioning the same issue repeatedly.

Similar as for usability testing [9], we also count the number of people that raised an issue. This is where using different note colors comes in handy as we can glance at a category and quickly get a sense of how many different people mention a certain issue

(Fig. 4b). By doing this, we are able to quantify our (mostly) qualitative findings. An opening statement such as “*most participants (16/20) explicitly said the prototype was easy to use*” will usually accompany a qualitative finding. Such statements allow us to shed light on and better ground our qualitative findings.

**Cleaning Up and Discarding the Affinity Wall.** After the affinity diagram has been duly documented, the room must be cleaned up. Besides taking the affinity wall and the (mostly) empty sheets of A3 paper down from the wall (Fig. 4d), the room needs to be arranged back to its original state (i.e., moving tables and chairs, and packing materials). Lining up the wall space with large sheets of paper at the start of the process for note clusters to form can greatly speed up the clean up process. Removing large sheets of paper is easier done and faster than manually removing 500 to 2500 notes one by one. Rolling up the affinity diagram to temporarily store it should be avoided, as the notes will tend to bend and might altogether fall [19]. Once the affinity wall has been taken down, and depending on internal practices regarding data handling and privacy, it is time to shred and discard the data (Fig. 4e). Depending on the final number of notes, cleaning up and discarding the wall can take up to two hours.

## 5 Discussion

### 5.1 Return on Investment and Impact

A recurring question within organizations when using affinity diagrams in interaction design evaluations is that of resources. Beyer and Holtzblatt [2] first collect data from 15–20 participants, producing 50 to 100 notes for each two-hour interview, for a total of 1500 notes. They then recommend having one person per 100 notes to build the affinity diagram in one day, for a total of 15 people. Due to the large number of participants (i.e., 10 to 24) and the resulting notes involved (i.e., 500 to 2500), our affinity diagramming process takes a team of two researchers between two and three weeks to collect the data (i.e., one week) and complete the analysis (i.e., one to two weeks). It takes us two to five days to individually write affinity notes, two to five days to build the affinity wall with the affinity team, and one or two days for an assigned scribe to document the wall. Holtzblatt et al. [13] describe two-person projects where it can take two to three weeks to gather requirements for participant numbers of less than ten. Therefore, we feel that our two-person team being able to analyze data from up to 24 people in the same amount of time is a good success indicator in terms of resources. Thanks to the mix of qualitative and quantitative analysis, plus the level of detail in our findings for each prototype evaluation, we have been able to easily justify assigning two people full-time to work on an affinity diagram.

Regarding impact, by systematically introducing affinity diagramming to our industrial research projects we were able to not only find existing UX and usability issues with our interaction designs, but also to identify and define new lines of research, requirements that had to be integrated to our designs, and new ideas that were filed as invention reports. While most resulting affinity diagram clusters covered issues with the interactive prototype that was currently being evaluated, for every project we had two

additional panels, one labeled ‘Future Research Areas/Other Topics’ and another one ‘Ideas/IPR’.

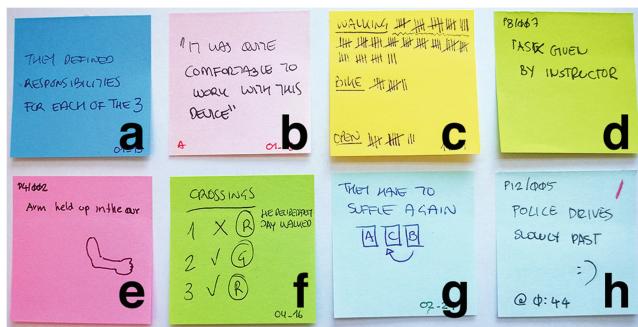
## 5.2 Number of Notes for Observations of Use and Interview

When creating affinity notes for interaction design studies, it is difficult to estimate the ratio of observation of use notes during the task to semi-structured interview notes. For a 45-min video consisting of 15 min of data for the task and 30 min for the semi-structured interview, one would perhaps expect a 1:2 ratio in the final number of notes as can be seen from Fig. 1b, where 20 and 40 notes were made respectively. However, we have come across unusually large numbers of notes from the observation part, sometimes even reaching 50 % of the total number of notes. The evaluation of Prototype E (Table 1) was related to groupware and explored general group formation using mobile phones for collaborative interactions. The fact that there were six participants trying out different strategies to form a group made it increasingly difficult to keep track of the overall situation. Tasks were taking place in parallel and there were many micro interactions happening. Although the task videos in this case were relatively short (i.e., 15 min), the complexity of the interaction data had a big impact on the final number of notes. Similarly, we roughly estimate that for a 45-min video it would take 90 min (or double the time) to generate affinity notes. However, in the case we are currently discussing, the amount of details in the interaction also had an effect on the overall note-taking time.

Our suggestion is for the note takers to start the interpretation sessions, go through the videos for the same two or three participants, and check how many notes each has made for the task and the semi-structured interview parts for each participant. Although there can be differences in the level of detail that each note taker captures, this coordination provides a reasonable estimate of whether someone is focusing on too much detail or being too general before all sessions have been interpreted.

## 5.3 Note Types and Leftover Notes

The most common affinity note types include observations of use (Fig. 5a), good participant quotes for use in publications (Fig. 5b), and design issues or ideas. As stated earlier, roughly 15 % of the notes are discarded (i.e., notes that are not included on the final affinity wall). Some of these notes are used to perform general counting (e.g., how many pedestrians and bicycles a participant encountered during the task) (Fig. 5c), to mark certain parts of the evaluation (e.g. “*Task given by [the] instructor*”) (Fig. 5d), to draw an arm posture (Fig. 5e), to personally indicate how well a task was performed (Fig. 5f), to draw different group formation strategies (Fig. 5g), or to record contextual information (e.g., “*Police drives slowly past:)*” (Fig. 5h). While note types 5c-5 h may at first seem superfluous, we argue for their importance from a holistic interaction design perspective as they help researchers thoroughly analyze micro-interactions, social interactions, and other contextual factors that may have an effect on the overall results.



**Fig. 5.** Leftover affinity notes. (a) Observations of use, (b) participant quote, (c) general counting, (d) personal separator, (e) drawing of body posture, (f) personal counting, (g) formation strategies, and (h) personal contextual observation.

Prototypes F, G and H (Table 1) have unusually large numbers of unused notes (23 %, 39 %, and 26 %). Based on the total number of affinity notes, these three projects could be considered average in size at 505, 1037, and 1276 notes. These three evaluations were conducted individually, which may cause note takers to make and write down more disconnected, random, and anecdotal observations. However, we do not believe this to be the sole source for the large number of discarded notes. We also ran out of time to go in more depth and further remove notes from the A3 sheets of paper to transfer them to the final affinity wall, as we spent three or four days to build these affinity walls.

## 6 Conclusion

By reflecting on a decade's experience using affinity diagramming across a number of projects, we have discussed how we have tailored the process for use in HCI and interaction design evaluations to four stages: *creating notes*, *clustering notes*, *walking the wall*, and *documentation*. Digital affinity diagrams can be especially convenient for situations when data are available in digital format (e.g., tweets, Facebook or YouTube comments), allowing single users to perform data coding, clustering, counting, and statistics in Excel, and are easy to transport. Despite existing attempts to augment affinity diagramming by making parts of the process digital, we have found that traditional paper affinity diagrams are better suited for collaborative analysis, they support building a cognitive model of the data by social and spatial awareness, and allow people to quickly transition from directly moving data around to having the full overview of the wall by simply taking a few steps away from or towards the wall. Providing such a flexible access to data in digital format, and in a way that is suitable for collaborative analysis would require the use of very large displays. In addition, we have embraced the affordances of paper by producing handwritten sticky notes, as we have not found alternative digital note types to offer clear advantages (i.e., speed, searchability). In tailoring Beyer and Holtzblatt's affinity diagramming process to provide better support when analyzing interactive prototype evaluations, we

have been able to better understand the context of use by checking the data sequentially, perform a micro-interaction analysis by creating and looking into detailed notes that might otherwise be discarded (e.g., counting, body postures, strategies), and to quantify qualitative findings at a glance using color-coded sticky notes.

**Acknowledgments.** A big shout out to Dima Aliakseyeu, Selene Mota, Marion Boberg, Hannu Korhonen, Jussi Holopainen, Tero Jokela, Akos Vetek, Jari Kangas and Deepak Akkil who co-built the affinity walls shown here. Thanks also to Jacob Buur, Robb Mitchell, and the anonymous reviewers for insightful comments on this paper.

## References

1. Benyon, D.: *Designing Interactive Systems: A Comprehensive Guide to HCI, UX and Interaction Design*. Pearson Education, Harlow (2013)
2. Beyer, H., Holtzblatt, K.: *Contextual Design: Defining Customer-Centered Systems*. Morgan Kaufmann, San Francisco (1998)
3. Buur, J., Soendergaard, A.: Video card game: an augmented environment for user centred design discussions. In: DARE 2000, pp. 63–69. ACM (2000)
4. Cox, D., Greenberg, S.: Supporting collaborative interpretation in distributed Groupware. In: CSCW 2000, pp. 289–298. ACM (2000)
5. Curtis, P., Heiserman, T., Jobusch, D., Notess, M., Webb, J.: Customer-focused design data in a large, multi-site organization. In: CHI 1999, pp. 608–615. ACM (1999)
6. Gaver, W.: Making spaces: how design workbooks work. In: CHI 2011, pp. 1551–1560. ACM (2011)
7. Geyer, F., Pfeil, U., Höchtl, A., Budzinski, J., Reiterer, H.: Designing reality-based interfaces for creative group work. In: C&C 2011, pp. 165–174. ACM (2011)
8. Geyer, F., Pfeil, U., Budzinski, J., Höchtl, A., Reiterer, H.: AffinityTable - a hybrid surface for supporting affinity diagramming. In: Campos, P., Graham, N., Jorge, J., Nunes, N., Palanque, P., Winckler, M. (eds.) *INTERACT 2011, Part III. LNCS*, vol. 6948, pp. 477–484. Springer, Heidelberg (2011)
9. Hanington, B.: *Universal Methods of Design: 100 Ways to Research Complex Problems, Develop Innovative Ideas, and Design Effective Solutions*. Rockport Publishers, Beverly (2012)
10. Harboe, G., Minke, J., Ilea, I., Huang, E.M.: Computer support for collaborative data analysis: augmenting paper affinity diagrams. In: CSCW 2012, pp. 1179–1182. ACM (2012)
11. Harboe, G., Doksam, G., Keller, L., Huang, E.M.: Two thousand points of interaction: augmenting paper notes for a distributed user experience. In: Lozano, M.D., Gallud, J.A., Tesoriero, R., Penichet, V.M.R. (eds.) *Distributed User Interfaces: Usability and Collaboration*, pp. 141–149. Springer, London (2013)
12. Hartson, R., Pyla, P.S.: *The UX Book: Process and Guidelines for Ensuring a Quality User Experience*. Morgan Kaufmann, Amsterdam (2012)
13. Holtzblatt, K., Wendell, J.B., Wood, S.: *Rapid Contextual Design*. Morgan Kaufmann, San Francisco (2005)
14. Jokela, T., Lucero, A.: MixedNotes: a digital tool to prepare physical notes for affinity diagramming. In: AcademicMindTrek 2014, pp. 3–6. ACM (2014)
15. Judge, T.K., Pyla, P.S., McCrickard, D.S., Harrison, S., Hartson, H.R.: Studying group decision making in affinity diagramming. Technical report TR-08-16, Computer Science, Virginia Tech (2008)

16. Kawakita, J.: The original KJ method. Kawakita Research Institute, Tokyo (1991)
17. Kildal, J., Lucero, A., Boberg, M.: Twisting touch: combining deformation and touch as input within the same interaction cycle on handheld devices. In: MobileHCI 2013, pp. 237–246. ACM (2013)
18. Klemmer, S.R., Newman, M.W., Farrell, R., Bilezikjian, M., Landay, J.A.: The designers' outpost: a tangible interface for collaborative web site. In: UIST 2001, pp. 1–10. ACM (2001)
19. Koh, B., Su, R.: Interactions for addressing shortcomings encountered when physically creating and manipulating large affinity diagrams. <http://www.benjaminkoh.com/AffinityTool.html>
20. Koskinen, I., Zimmerman, J., Binder, T., Redström, J., Wensveen, S.: Design Research Through Practice: From the Lab, Field, and Showroom. Morgan Kaufmann, San Francisco (2011)
21. Kuniavsky, M.: Observing the User Experience: A Practitioner's Guide to User Research. Morgan Kaufmann, San Francisco (2003)
22. Lucero, A., Lashina, T., Diederiks, E. and Mattelmäki, T.: How probes inform and influence the design process. In: DPPI 2007, pp. 377–391. ACM (2007)
23. Lucero, A., Boberg, M., Uusitalo, S.: Image space: capturing, sharing and contextualizing personal pictures in a simple and playful way. In: ACE 2009, pp. 215–222. ACM (2009)
24. Lucero, A.: Framing, aligning, paradoxing, abstracting, and directing: how design mood boards work. In: DIS 2012, pp. 438–447. ACM (2012)
25. Lucero, A., Jones, M., Jokela, T., Robinson, S.: Mobile collocated interactions: taking an offline break together. *Interactions* **20**(2), 26–32 (2013)
26. Lucero, A., Vetek, A.: NotifEye: using interactive glasses to deal with notifications while walking in public. In: ACE 2014, Article 17, 10 p. ACM (2014)
27. Luff, P., Heath, C.: Mobility in collaboration. In: CSCW 1998, pp. 305–314. ACM (1998)
28. Luff, P., Heath, C., Norrie, M., Signer, B., Herdman, P.: Only touching the surface: creating affinities between digital content and paper. In: CSCW 2004, pp. 523–532. ACM (2004)
29. Mayhew, D.J.: The Usability Engineering Lifecycle: A Practitioner's Guide to User Interface Design. Morgan Kaufmann, San Francisco (1999)
30. Miura, M., Sugihara, T., Kunifugi, S.: GKJ: Group KJ method support system utilizing digital pens. *IEICE Trans. Inf. Syst.* **94**(3), 456–464 (2011)
31. Nunes, M., Greenberg, S., Neustaedter, C.: Sharing digital photographs in the home through physical mementos, souvenirs, and keepsakes. In: DIS 2008, pp. 250–260. ACM (2008)
32. Wallace, J., McCarthy, J., Wright, P.C., Olivier, P.: Making design probes work. In: CHI 2013, pp. 3441–3450. ACM (2013)

# What Users Prefer and Why: A User Study on Effective Presentation Styles of Opinion Summarization

Xiaojun Yuan<sup>1</sup>(✉), Ning Sa<sup>1</sup>, Grace Begany<sup>1</sup>, and Huahai Yang<sup>2</sup>

<sup>1</sup> College of Computing and Information,  
University at Albany, State University of New York, Albany, USA

{xyuan, nsa, gbegany}@albany.edu

<sup>2</sup> Juji, Inc., Saratoga, USA  
hyang@juji-inc.com

**Abstract.** Opinion Summarization research addresses how to help people in making appropriate decisions in an effective way. This paper aims to help users in their decision-making by providing them effective opinion presentation styles. We carried out two phases of experiments to systematically compare usefulness of different types of opinion summarization techniques. In the first crowd-sourced study, we recruited 46 turkers to generate high quality summary information. This first phase generated four styles of summaries: Tag Clouds, Aspect Oriented Sentiments, Paragraph Summary and Group Sample. In the follow-up second phase, 34 participants tested the four styles in a card sorting experiment. Each participant was given 32 cards with 8 per presentation styles and completed the task of grouping the cards into five categories in terms of the usefulness of the cards. Results indicated that participants preferred Aspect Oriented Sentiments the most and Tag cloud the least. Implications and hypotheses are discussed.

**Keywords:** Text summarization · Consumer decision making · User studies · User interface design

## 1 Introduction

The widespread use of the Internet in many aspects of human activities has resulted in an abundance of publicly-accessible opinions. People can find opinions on a variety of topics in venues such as Twitter, Weibo, forums, e-commerce sites and specialized opinion-hosting sites such as Yelp. While most of these opinions are intended to be helpful for others, the sheer quantity of them often makes most of the opinions underutilized, as the information overload overwhelms many potential consumers. For example, Amazon.com has more than 18,000 reviews for Kindle Fire, a single product alone. Summarizing these reviews in some concise form could bring enormous benefits to consumers and business alike. Not surprisingly, research on opinion summarization is gaining increased attention [14, 18, 19, 22, 24]. However, most of the research emphasizes technical advances in underlying algorithms, while paying less attention to the presentation of the results, which is the focus of this work. Correspondingly,

evaluation of opinion summarization research is normally based on certain notions of precision and recall calculation commonly used in information retrieval [28] and data mining [29]. Studies have only begun to investigate the effectiveness of opinion summarization in term of usability (e.g. [31]). Such studies focus on testing the newly-proposed techniques. A systematic comparison of the usefulness of different types of opinion summarization is still lacking. This paper reports our effort in addressing this deficiency.

One major difficulty with studying the effectiveness of opinion summarization is a confounding effect between content effectiveness and presentation effectiveness. It is often not clear whether a technique's empirical superiority can be attributed to its superior text analytics quality or its effective information presentation style. We plan to isolate the two factors and focus on studying the effect of presentation styles. This goal is achieved by using human-generated summarization as the content, so as to ensure the content has consistent high quality regardless of the presentation styles. We can then vary the presentation styles of the summaries to investigate their effect on the usefulness ratings of the summaries. Any differences found between the usefulness ratings of the summaries can be safely attributed to the differences in presentation styles. We identified four types of presentation styles of opinion summarization through a crowd-sourcing study on Amazon Mechanical Turk, and then conducted a lab user-centered experiment to compare the effectiveness of the four styles.

## 2 Previous Work

Although not abundant, studies investigating the effectiveness of opinion summarization from a perspective of both usability and user preference are emerging. Several recent studies explore feedback from users regarding their preferences for certain opinion summarization styles and approaches. Most recently, Qazi et al. [26] addressed a gap in existing studies examining the determination of useful opinion review types from customers and designers perspectives. Specifically, the researchers used the Technology Acceptance Model (TAM) as a lens to analyze users' perceptions toward different opinion review types and online review systems. The study, a pilot study, focused on three review types which are related to perceived usefulness, perceived ease of use, and behavioral intention: A (regular), B (comparative), and C (suggestive). Suggestive reviews, the speech acts which are used to direct someone to do something in the form of a suggestion, were newly identified by the researchers as a third innovative review type. To examine user perspectives, researchers used a closed card sorting approach to analyze reviews from Amazon, blogs, and a self-deployed website. The results of their work indicated that the review types play a significant role in developing user perception regarding a new product or system, with suggestive reviews more significant for both customers and designers to find more usefulness that ultimately improves their satisfaction level.

Further, in another work [31], researchers conducted a user study of a review summarization interface they created called "Review Spotlight." Review Spotlight is based on a tag cloud and uses adjective-noun word pairs to provide an overview of online restaurant reviews. Findings indicated that study participants could form detailed

impressions about restaurants and make faster decisions between two options with Review Spotlight versus traditional review webpages. In a large-scale, comprehensive human evaluation of three opinion-based summarization models – Sentiment Match (SM), Sentiment Match + Aspect Coverage (SMAC), and Sentiment-Aspect Match (SAM) – Lerman, Blair-Goldensohn and McDonald [15] found that users have a strong preference for sentiment-informed summaries over simple, non-sentiment baselines. This finding reinforces the usefulness of modeling sentiments and aspects in opinion summarization. In another study, Lerman and McDonald [16] investigated contrastive versus single-product summarization of consumer electronics and found a significant improvement in the usefulness of contrastive summaries versus summaries generated by single-product opinion summarizers. To find out which visual properties influence people viewing tag clouds, Bateman, Gutwin and Nacenta [2] conducted an exploratory study that asked participants to select tags from clouds that manipulated nine visual properties (font size, tag area, number of characters, tag width, font weight, color, intensity, and number of pixels). Participants were asked to choose tags they felt were “visually important” and results were used to determine which visual properties most captured people’s attention. Study results indicated that font size and font weight have stronger effects than intensity, number of characters or tag area. However, when several visual properties were manipulated at one time, no one visual property stood out among the others. Carenini, Ng and Pauls [4] also employed a user study as part of their wider comparison of a sentence extraction-based versus a language generation-based summarizer for summarizing evaluative text. In their quantitative data analysis, the researchers found that both approaches performed equally well. Qualitative data analysis also indicated that both approaches performed well, however, for different, complementary reasons. In a related work, Carenini, Ng and Pauls [5] examined the use of an interactive multimedia interface, called “Treemaps,” for summarizing evaluative text of online reviews of consumer electronics. Treemaps presents the opinion summarizations as an interactive visualization along with a natural language summary. Results of their user study showed that participants were generally satisfied with the interface and found the Treemap summarization approach intuitive and informative.

In more recent work, [12] researchers presented a novel interactive visual text analytic system called, “OpinionBlocks.” OpinionBlocks had two key design goals: (1) automated creation of an aspect-based visual summary to support users’ real-world opinion and analysis tasks, and (2) support of user corrections of system text analytic errors to improve system quality over time. To demonstrate OpinionBlock’s success in addressing the design goals, researchers employed two crowd-sourced studies on Amazon Mechanical Turk. According to their results, over 70 % of users successfully accomplished non-trivial opinion analysis tasks using OpinionBlocks. Additionally, the study revealed that users are not only willing to use OpinionBlocks to correct text classification mistakes, but that their corrections also produce high quality results. For example, study participants successfully identified numerous errors and their aggregated corrections achieved 89 % accuracy.

Additionally, Duan et al. [7], introduced the opinion mining system, “VISA” (VIvisual Sentiment Analysis), (derived from an earlier system called TIARA). The VISA system employs a novel sentiment data model to support finer-grained sentiment analysis, at the core of which is the “sentiment tuple,” composed of four

elements: feature, aspect, opinion, and polarity. Researchers conducted a user experiment to explore how efficiently people could learn to use VISA and demonstrate its effectiveness. Study results indicated that VISA performed significantly better than the two comparison tools (TripAdvisor and a text edit tool) due to its key features, namely mash-up visualizations and rich interaction features.

In the current study, investigation of user perspectives on opinion summarization styles is taken further with the evaluation and comparison of four distinct, popular summarization styles focused on textual opinions; namely, Tag Clouds, Aspect-Oriented Sentiments, Paragraph Summaries, and Group Samples.

Following we introduce the four presentation styles used in our study, the methodology, results of the experiment, discussion and conclusions.

### 3 Opinion Summarization Presentation Styles

Some opinion hosting sites allow opinion writers to give numerical ratings in addition to the textual opinions. Since the visualization of numerical values is a well-studied problem, we focus instead on the summarization of textual opinions. Similarly, we do not compare visualization systems that emphasize statistics rather than the textual content of the text collections (e.g. [6]). Opinion summarizations studied here are of the kind that could potentially be used in place of the full documents.

Based on our survey of the literature, we have categorized the presentation styles of such opinion summarization into four major types.

#### 3.1 Tag Clouds (TAG)

Tag clouds are perhaps the most popular form of summarization on the Internet today [3]. This type of text presentation has also been used extensively in research (e.g. [25, 28]). They consist of enumerations of the most common words or phrases in a collection of opinions, laid out on a 2D canvas. The size of the words or phrases often indicates how frequently they were mentioned. The larger the word or phrase, the more frequently it received mentions. The effect of various visual features of the tag clouds on their effectiveness have been investigated [2], but the comparison with other styles of summarization has yet to be done. See Fig. 1.

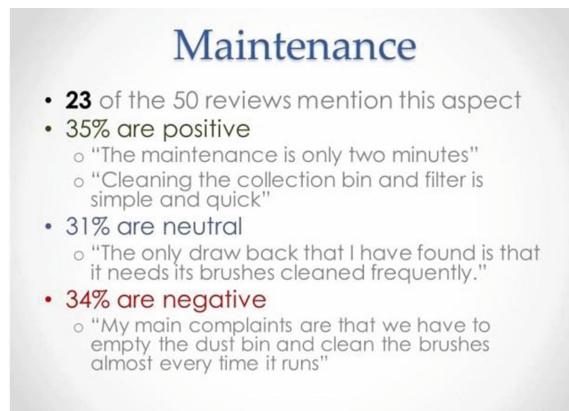
#### 3.2 Aspect Oriented Sentiments (ASP)

Aspect oriented sentiment summarization is an active area of research in text mining [11, 13, 14, 19, 23]. In this approach, some important aspects or topics (also known as features) of opinions are extracted from an opinion collection. Sentiment orientation of the text snippets containing the aspects are then estimated and summary statistics reported. A typical summarization for one aspect might look like this: *for a collection of reviews on Kindle Fire, “screen” is identified as an aspect, and 100 text snippets in the collection are found to be about this aspect, 60 of them have positive sentiment, 30*



**Fig. 1.** Tag clouds

of them are negative, and the rest are neutral. Representative text snippets for each sentiment orientation may also be provided. See Fig. 2.



**Fig. 2.** Aspect oriented sentiments

### 3.3 Paragraph Summaries (PRG)

Automatic text summarization systems traditionally produce short passages of text as summary [10, 27]. The summarization is called extractive when the sentences are selected from original documents; abstractive when the sentences are generated by the system [8]. Regardless of the approach, the output could be a readable abstract that resembles what humans would write for generic purposes in order to emphasize intrinsic properties such as fluency and coverage [21]. See Fig. 3.

### Summary Paragraph 8

The Roomba is extremely effective at cleaning up dog/cat hair and dirt on both carpets and bare floor, transitioning easily between the two. It does, however, require frequent cleaning of the brushes and emptying of the bin, sometimes in the middle of cleaning. Long hair tends to get wrapped around the brushes and needs to be removed. The Roomba avoids sucking up cords for the most part, but can end up pulling in small cables if they're thin enough. The overall durability of the Roomba is questionable, as the lifespan of the unit only seems to average a year or two. Support from iRobot is difficult to get and often unhelpful. The unit can occasionally get stuck under or behind furniture like couches or beds and will need to be rescued, but this is rare. In very large living areas it can occasionally get lost. Most owners seem very happy with the cleaning job it does despite these issues, and consider it valuable.

**Fig. 3.** Paragraph summary

### 3.4 Group Samples (GRP)

Clustering algorithms are typically used in post-query user interfaces as a way of organizing retrieved documents [9]. It has also been used in summarization, where similar documents are grouped together and the representatives of the groups are displayed [1, 20]. This approach has shown to be effective in interactive information retrieval [17, 30]. See Fig. 4.

#### Group C contains 9 similar reviews

*The most representative review:*

**Careful!**

Beware! No roomba works on dark surfaces. I had a roomba and then bought a new living room rug -- a patterned oriental rug that has a lot of navy blue and black in it. I was told by the i-robot company that the reason my roomba is not working is that none of them work on dark surfaces. I have not seen that warning anywhere in the product literature -- because it was more than 30 days since I purchased my roomba they refused to refund my money. I wonder what else they are not being forthright about?

**Fig. 4.** Group samples

### 3.5 Summary of the Four Presentation Styles

Tag clouds may perform most differently from the other three presentation styles. Because Tag clouds do not show full sentences, they lack the context needed to make accurate judgments about the value of the information. On the other hand, Tag clouds

have the best support for fast perception as they package the most important words or phrases efficiently in space, and visually highlight their relative importance. Aspect oriented sentiments are similar to Tag clouds in that they lack a prose structure. On the other hand, they are also similar to Paragraph summaries because they include text snippets that are reader friendly and provide the context missing in Tag clouds. Paragraph summaries are close to Aspect oriented sentiments because they may cover similar amounts of information, as paragraphs often list pros and cons in a form that resembles aspect oriented sentiments. However, the prose structure in a well-written paragraph summary affords deep analysis and accurate assessment of the context. Group samples are similar to paragraph summaries in form. However, unlike paragraph summaries that are written anew, group samples are directly drawn from the original document collection, and retain the most amount of contextual information.

## 4 Research Questions

We hypothesize that humans respond differently to different presentation styles of opinion summaries, and some styles would be more effective in terms of human acceptance.

In this two-phase study, we are interested in investigating the following research questions:

1. Will users prefer (or not prefer) a particular opinion summarization style in making judgments about product reviews?
2. What are the reasons that users may prefer (or not to prefer) a particular opinion summarization style in making judgments about product reviews?

### 4.1 Phase I: Crowd Sourcing Opinion Summarization

As mentioned earlier, in order to study the effect of presentation styles alone, we want to ensure the consistently high quality of the summaries. We achieved the goal through leveraging the wisdom of the crowd. Essentially, we elicited four styles of the opinion summaries with the help of Amazon's Mechanical Turk.

**Procedure.** First, we collected the top 50 reviews for one model of the iRoomba cleaning robot from Amazon. We chose this collection of opinion text because of the relative novelty of the product and the ability for the general public to relate to it. Using a within-subject design, we recruited 46 turkers located in the USA to answer a survey we developed to gather information for generating the opinion summaries from the text collection. In the survey, turkers were first directed to the raw text of the 50 reviews, and asked to read them in full. Then, questions about the reviews were asked. These questions were directly mapped from the information need for the four presentation styles of summaries. All questions were mandatory and were individually validated to ensure the quality of the answers. On average, turkers spent 56.3 min on the survey, and each was paid 4 US dollars.

**Generating Opinion Summaries Using Turkers.** For Tag clouds, turkers were asked to list five short phrases to characterize the cleaning robot. They were also asked to estimate what percentage of the 50 reviews had opinions consistent with each phrase. The phrases turkers came up with were remarkably consistent and converged to 38 phrases (phrases with minor variations were grouped as one). All 38 phrases were used in the subsequent lab study. The average percentages of turkers' estimations were used in the subsequent lab study to determine the font size of the phrases. A total of 8 Tag clouds were drawn by hand, with each cloud containing 4 or 5 phrases.

Turkers were asked to list three important aspects of the product according to the reviews they read. For each aspect, they were asked to give an estimate of how many reviews mentioned the aspect, as well as the estimated percentage for positive, neutral and negative sentiment towards the aspect. The top 8 most-frequently listed aspects were used in the subsequent study. Again, the averages of the estimations were used in the display of the aspects.

Each turker was also asked to write a summary of all the reviews so that "consumers who read your summary can make an informed decision on the product, without having to read the reviews themselves". Among the 46 summaries, the top 8 most readable summaries, as agreed by two judges, were used in the lab study.

We asked turkers to identify similar reviews and group them together. They were required to list 3 groups of similar reviews.

When two reviews appeared in the same group once, their similarity measure increase by one. This way, we were able to generate a similarity matrix among the 50 reviews.

Using the matrix as input, we used a hierarchical clustering algorithm to cluster the 50 reviews. Four clusters produced the optimal fit, and the four cluster prototypes were used in the lab study as group samples. In addition, for each cluster, the closest summary to the prototype was also selected, so that there were 8 group samples in total.

## 4.2 Phase II: Comparing Presentation Styles of Summary Using Card Sorting Technique

The goal of this phase was to compare the four presentation styles of opinion summary in a consumer decision making context to respond to the two research questions.

**Experiment Design.** The comparison of the opinion summaries was conducted as a lab card-sorting task. For each of the four presentation styles (experiment conditions), eight opinion summaries were prepared according to the procedure described in the previous section.

Each opinion summary was put on a single image of  $960 \times 720$  pixels. Figures 1, 2, 3 and 4 show a sample display for each of the conditions. In total,  $4*8 = 32$  image items were placed in the preparation bin in a random order.

The lab experiment was a within-subject design. Each participant's task was to take all 32 image items and place each of them in one of the five category bins. These category bins were defined as "Not at all useful", "Somewhat useful", "Useful", "Very useful" and "Extremely useful". Participants were told to ignore the card order within

each category box. Essentially, we asked participants to give a usefulness rating for each opinion summary. We use such a card-sorting setting in order to record participants' thought process, as they were asked to think-aloud while placing the cards.

**Participants.** Thirty-four participants were recruited from University at Albany, half of them were males. All of the participants stated that they read online reviews regularly for making purchase decisions.

**Procedure.** Each participant was tested individually in a human computer interaction lab in a University campus in the USA. The subjects first filled out a consent form. Next, the subjects completed an entry questionnaire. The participants were then directed to <http://websort.net/> to do the card sorting. After they completed the experiment, they were asked to answer several questions regarding the four presentation styles and their thoughts about the experiment. The whole experimental process was logged by Techsmith Morae 3 software.

## 5 Content Analysis Scheme

To address the earlier mentioned research question(What are the reasons that users may prefer (or not prefer) a particular opinion summarization style in making judgments about product reviews?) we employed a qualitative content analysis using an open coding approach to analyze the exit interview data of the Phase II experiment. The content analysis began with a comprehensive read-through and evaluation of all 34 participants' interview transcripts by each of the first three authors, the primary investigator and two doctoral students. In terms of the initial review and several discussions between the authors, a number of themes emerged from the interviews that pertained to the reasons of preference towards the presentation styles. Themes included: Comprehensiveness (Comprehensiveness of Information), Time (Time required to read the summary), Organization/categorization (Organization/categorization of the summary's content), Length/Amount (Length/amount of information), Appearance (Appearance of summary content), and Ease of use (Ease of use of summarization style). A coding scheme was designed according to these themes and is shown in the Table 1.

The unit of analysis for the open coding was the individual interview document. Each of the 34 interview documents were independently coded by the three researchers and data collected in an Excel spreadsheet. The average pairwise percent agreement among the 3 coders is 0.81. Along with each code, snippets of supporting text were extracted from the interview data.

## 6 Results

### 6.1 User Perception of the Presentation Styles

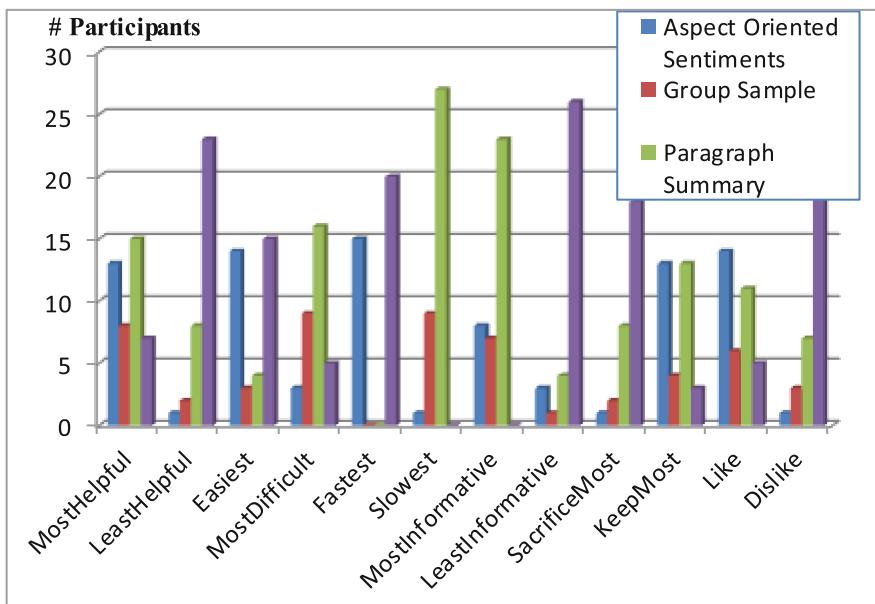
As initially noted, the goal of this paper is to better understand the relationship between opinion summarization styles and user decision judgment and the underlying reasons.

**Table 1.** Coding scheme

Code	Reasons
CMPP CMPNCMPU	Comprehensiveness (Comprehensiveness of Information) ( <b>CMP</b> )
TMEP TMEN TMEU	Time (Time required to read the summary) ( <b>TME</b> )
ORGP ORGN ORGU	Organization/categorization (Organization/categorization of the summary's content) ( <b>ORG</b> )
LENP LENN LENU	Length/Amount (Length/amount of information) ( <b>LEN</b> )
APPP APPN APPU	Appearance (Appearance of summary content) ( <b>APP</b> )
E2UP E2UN E2UU	Ease of use (Ease of use of summarization style) ( <b>E2U</b> )

The results are aligned with this goal and are based on a qualitative content analysis of the participant exit interview data.

In the exit interview, the participants answered questions relevant to the four presentation styles, including, (1) most helpful/least helpful, (2) easiest/most difficult, (3) fastest/slowest, (4) most informative/least informative, (5) sacrifice/keep the most, and (6) like/dislike. The six questions were the basis for the measurement of user perception of the four presentation styles. The results are displayed in Fig. 5.



**Fig. 5.** User perception of the four presentation styles (y axis is the number of participants who voted for the style)

Participant responses also included their opinions on the usefulness of the summary in general and strategies they used to make the decision, as well as suggestions on system features they best liked, disliked, and thought could be added in the future.

Although “think-aloud” data and additional computer log data was collected using the Morae software, the current paper focuses exclusively on results from the exit interview data.

As can be seen from Fig. 5, participants felt that paragraph summary was the most helpful, and the most informative presentation style while it was also the most difficult and the slowest one. Tag clouds were reported to be the easiest and the fastest to use, but accordingly they were the least informative and the least helpful style and the one disliked by most of the participants. On the other hand, though Aspect Oriented Sentiments didn’t get the most votes, they were found to be generally helpful, easy to use, fast, and informative and the participants liked them the most. Group samples were relatively less helpful, more difficult, and slower.

## 6.2 Users’ Opinion on the Presentation Styles

We were interested in finding out the major reasons influencing users’ preference of the various opinion summarization styles. It appears that “Comprehensiveness”, “Organization/Categorization” and “Ease of use” are equally important key reasons affecting users’ preference of the presentation styles. The “Appearance”, “Time”, and “Length/Amount” were found to be less important reasons to the participants. Table 2 shows the distribution of reasons coded across participants. To generate the table, each unique code instance was counted for each participant.

**Table 2.** Major reasons affecting user preference of the four presentation styles

Code	Reasons	No. responses
CMPP; CMPN; CMPPU	CMP	33
ORGP; ORGN; ORGU	ORG	32
E2UP; E2UN; E2UU	E2U	30
APPP; APPN; APPU	APP	27
TMEP; TMEN; TMEU	TME	26
LENP; LENN; LENU	LEN	26

We further investigated the distribution of the identified reasons across the four presentation styles. Table 3 shows the top 5 reasons for each style. Tag clouds received positive comments on “Appearance” and “Time” and a nearly equal amount of positive and negative comments on “ease of use.” However, most of the participants agreed that they were negative regarding “comprehensiveness.” This finding can explain the above-mentioned finding that participants disliked tag clouds the most. As mentioned by participants, “They don’t, they don’t have details.” All of the top 5 reasons related to aspect oriented sentiments were positive and covered all of the main reason categories except “time.” The participants liked them because they “contain negative and neutral

and the positive opinions,” were “very convenient or easy to read,” and “very clear, brief,” among other reasons. This finding correlates with the findings in Fig. 5 and also explains why the participants liked aspect oriented sentiments the most. Most participants found the paragraph summary good in terms of “comprehensiveness” and some liked its organization and found the formatting was “...what’s most normal for me.” But on the other hand, the paragraph summary received negative comments regarding “time” and “length.” The participants found it “too long to read” and they needed to “Spend time reading it.” Compared with the previous three styles, the group sample received far fewer comments. Some of the participants mentioned that it was good in terms of “comprehensiveness” and “organization,” but, some didn’t like it because of the “length,” “ease of use,” and “organization.” Typical user comments can be found in Table 3.

Unsurprisingly, in both paragraph summary and the group sample styles, comprehensiveness was the predominant reason in users’ preference decision. Specifically, participants claimed that the paragraph summary is “getting a lot of information in a fairly simple package” and the group sample helps them “imagine what if I had that product.”

For tag clouds, though the participants liked them because they were “much faster” and “the font size was there for the words,” they agreed that “They don’t, they don’t give details” and were negative regarding “comprehensiveness.” On the contrary, the paragraph summary was long and time-consuming, but most of the participants found it provided “a lot of information in a fairly simple package” and was positive in its “comprehensiveness.” Overall, aspect oriented sentiments were the best among the four styles. They were brief and, at the same time, comprehensive. The participants found them easy to use and liked their appearance and organization.

## 7 Discussion

In this paper, we were interested in discovering users’ preferences and the reasons affecting their preferences of the representation styles in an opinion summarization card-sorting task.

Results demonstrate that: (1) Aspect oriented sentiments are the most preferred presentation style; (2) comprehensiveness, time, organization/categorization, length/amount, appearance, and ease of use are the major reasons impacting users’ preferences of a presentation style in making decisions in a product review task.

Our results supported the finding reported in [3] in that our participants disliked the tag clouds the most. As [3] pointed out, “tags are useful for grouping articles into broad categories, but less effective in indicating the particular content of an article” In our study, the participants acknowledged that tag clouds were the easiest to use because “We can construe what the product was like in very short time,” and the fastest to use because “It’s very fast; the tag clouds was much faster.” But, in making the decision on the usefulness of the presentation styles, their first priority was the comprehensiveness of the information in the summary presentation. As mentioned by participants, they understood people liked the tag clouds because of the “font size” and “color,” but, they disliked them because they don’t “give details” and “drive my decision making.” This

**Table 3.** Reasons affecting user preference per presentation style

Styles	Code	Reasons	No.	User comments
TAG	CMPN	CMP	27	They don't, they don't give details
	TMEP	TME	14	It's very fast; the tag clouds was much faster
	APPP	APP	13	I can know how many people like it, because the font size was there for the words
	E2UN	E2U	12	It was like I couldn't do anything with it. It was like it didn't seem to drive the decision, it wouldn't drive my decision making
	E2UP	E2U	11	We can construe what the product was like in very short time
ASP	ORGP	ORG	23	It contains negative and neutral and the positive opinions
	APPP	APP	15	Because color coding and numbers
	E2UP	E2U	15	They're very convenient or easy to read
	CMPP	CMP	10	Percentages mean a lot for people to review all these products, and there's examples, a lot of examples on it
	LENP	LEN	7	It was very clear, brief... concise information
PRG	CMPP	CMP	24	That's getting a lot of information in a fairly simple package
	TMEN	TME	20	Spend time reading it
	LENN	LEN	16	Too long to read
	ORGP	ORG	12	I enjoy the formatting and that's what's most normal for me
	E2UN & ORGN	E2U	10	... it's a lot of effort for little result sometimes. It's a tossup; It's because of so much text and even if you read it, it is not organized sometimes...
GRP	CMPP	CMP	10	Yeah the personal story with the rich information helps me to imagine what if I had that product
	ORGP	ORG	8	I'm seeing the whole review versus just one person's interpretation of all of the reviews
	LENN	LEN	7	In those group samples, that I think was too large is the content was too large
	E2UN	E2U	5	because they tended to repeat the same information
	ORGN	ORG	5	It was not categorized properly. And if it isn't organized properly, it's confusing

finding raised an important issue here for the design of information systems: How can the user interface balance the need for comprehensiveness of information and the need to provide key features enabling users to quickly grasp the desired information? On the other hand, [3] reported that, compared with human-assigned tags, automated tagging produces “more focused, topical clusters.” The tag clouds in our study were generated

by the turkers. As a result, the comprehensiveness of automatically generated tags might be our future research direction.

It was interesting to learn that participants liked the aspect oriented sentiments the most. Organization/categorization is the most critical reason in users' decision making relevant to this presentation style. Most importantly, they liked them because they contain "negative and neutral, and the positive opinions," "color," "number" and "percentages."

Our results indicated that there may be a relationship between consumers' information needs and their preference of an opinion summary presentation style. With regards to the opinion summary of a cleaning robot, it is within expectation that consumers may want to look for information about system usability, performance, and reliability. This factor may contribute to the finding that participants prefer aspect oriented sentiments, not tag clouds.

It can also be noticed that there exist biases potentially introduced by manually generated summary presentations. Our summarizations were generated by human turkers, but, this generation could have been influenced by the instructions and contents distributed by the researchers.

Results of this study have practical implications for developers of text summarization. A few design considerations for improving usability and user experiences emerged based on participant responses and our observations. First, a deeper understanding of users' information behavior and their information needs in using information systems supporting consumer decision-making is important. In this study, we made a step towards this understanding in using turkers to generate the summary reviews. Second, after having identified key features in the consumer decision-making system, a good design should well balance the number of the features and the amount of information provided in the interface. Third, an appropriate and comprehensive organization/categorization scheme should be selected in terms of targeted user groups and task and design considerations. Many participants expressed their opinions about the importance of organization/categorization. We feel it should be given greater attention in the design process in the future experiments.

## 8 Conclusion

This paper reports a study comparing the effectiveness of four major styles of opinion summarization in a consumer purchase decision scenario. The study leverages the power of the crowd to bypass issues of text mining quality in order to reach more meaningful conclusions.

Our next step is to design and implement an experimental system based on the findings of this study. Such an experimental system will provide customers with a better view in the system interface. Additionally, the experimental system will be compared with our baseline system in a user-centered lab experiment to test its effectiveness and efficiency. Our goal is to contribute to improving the user experience and usability of information systems that support consumer decision-making.

As a lab-based user-centered study, limitations exist. In this experiment, the generalizability of the findings was restricted by the limited types of tasks, the number of

topics, and the sample pool. Additionally, the coding scheme we generated is a simple, initial one. Deeper, more fine-tuned coding and analysis could be applied to the data in a subsequent analysis. Despite the limitations, the results of this type of research will have implications for the design of information systems that support consumer decision-making.

## References

1. Ando, R., Boguraev, B., Byrd, R., Neff, M.: Multi-document summarization by visualizing topical content. In: Proceedings of the 2000 NAACL-ANLP Workshop on Automatic Summarization, pp. 79–98 (2000)
2. Bateman, S., Gutwin, C., Nacenta, M.: Seeing things in the clouds: the effect of visual features on tag cloud selections. In: Proceedings of the Nineteenth ACM Conference on Hypertext and Hypermedia, pp. 193–202. ACM (2008)
3. Brooks, C., Montanez, N.: Improved annotation of the blogosphere via autotagging and hierarchical clustering. In: Proceedings of the 15th International Conference on World Wide Web, pp. 625–632 (2006)
4. Carenini, G., Ng, R., Pauls, A.: Multi-document summarization of evaluative text. In: Proceedings of the European Chapter of the Association for Computational Linguistics (EACL), pp. 305–312 (2006)
5. Carenini, G., Ng, R., Pauls, A.: Interactive multimedia summaries of evaluative text. In: Proceedings of IUI 2006 of the Association for Computing Machinery, pp. 1–8 (2006)
6. Chen, C., Ibekwe-SanJuan, F., SanJuan, E., Weaver, C.: Visual analysis of conflicting opinions. In: 2006 IEEE Symposium On Visual Analytics Science And Technology, pp. 59–66 (2006)
7. Duan, D., Qian, W., Pan, S., Shi, L., Lin, C.: VISA: a visual sentiment analysis system. In: Proceedings of VINCI 2012, Hangzhou, China (2012)
8. Erkan, G., Radev, D.: Lexrank: graph-based lexical centrality as salience in text summarization. J. Artif. Intell. Res. (JAIR) **22**, 457–479 (2004)
9. Hearst, M., Pedersen, J.: Reexamining the cluster hypothesis: scatter/gather on retrieval results. In: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 76–84 (1996)
10. Hovy, E., Lin, C.: Automated text summarization and the summarist system. In: Proceedings of a Workshop on Held at Baltimore, Maryland, 13–15 October 1998, pp. 197–214 (1998)
11. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 168–177 (2004)
12. Hu, M., Yang, H., Zhou, M.X., Gou, L., Li, Y., Haber, E.: OpinionBlocks: a crowd-powered, self-improving interactive visual analytic system for understanding opinion text. In: Kotzé, P., Marsden, G., Lindgaard, G., Wesson, J., Winckler, M. (eds.) INTERACT 2013, Part II. LNCS, vol. 8118, pp. 116–134. Springer, Heidelberg (2013)
13. Jo, Y., Oh, A.: Aspect and sentiment unification model for online review analysis. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, pp. 815–824 (2011)
14. Ku, L., Liang, Y., Chen, H.: Opinion extraction, summarization and tracking in news and blog corpora. In: Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs (2006)

15. Lerman, K., Blair-Goldensohn, S., McDonald, R.: Sentiment summarization: evaluating and learning user preferences. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, pp. 514–522 (2009)
16. Lerman, K., McDonald, R.: Contrastive summarization: an experiment with consumer reviews. In: Proceedings of NAACL HLT of the Association for Computational Linguistics, pp. 113–116 (2009)
17. Leuski, A.: Evaluating document clustering for interactive information retrieval. In: Proceedings of the Tenth International Conference on Information and Knowledge Management, pp. 33–40 (2001)
18. Lu, Y., Zhai, C.: Opinion integration through semi-supervised topic modeling. In: Proceedings of the 17th International Conference on World Wide Web, pp. 121–130 (2008)
19. Lu, Y., Zhai, C., Sundaresan, N.: Rated aspect summarization of short comments. In: Proceedings of the 18th international conference on World wide web, pp. 131–140 (2009)
20. Maña-López, M., De Buenaga, M., Gómez-Hidalgo, J.: Multidocument summarization: an added value to clustering in interactive retrieval. ACM Trans. Inf. Syst. (TOIS) **22**(2), 215–241 (2004)
21. Mani, I., House, D., Klein, G., Hirschman, L., Firmin, T., Sundheim, B.: The tipster summac text summarization evaluation. In: Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics, pp. 77–85 (1999)
22. Mei, Q., Ling, X., Wondra, M., Su, H., Zhai, C.: Topic sentiment mixture: modeling facets and opinions in weblogs. In: Proceedings of the 16th International Conference on World Wide Web, pp. 171–180 (2007)
23. Mukherjee, A., Liu, B.: Aspect extraction through semi-supervised modeling. In: Proceedings of the 50<sup>th</sup> Annual Meeting of the Association for Computational Linguistics, pp. 339–348 (2012)
24. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends Inf. Retrieval **2**(1–2), 1–135 (2008)
25. Potthast, M., Becker, S.: Opinion summarization of web comments. In: Gurrin, C., He, Y., Kazai, G., Kruschwitz, U., Little, S., Roelleke, T., Rüger, S., van Rijsbergen, K. (eds.) ECIR 2010. LNCS, vol. 5993, pp. 668–669. Springer, Heidelberg (2010)
26. Qazi, A., Raj, R.G., Tahir, M., Waheed, M., Khan, S.U.R., Abraham, A.: A preliminary investigation of user perception and behavioral intention for different review types: customers and designers perspective. Sci. World J. **2014**, 1–8 (2014)
27. Radev, D., McKeown, K.: Generating natural language summaries from multiple on-line sources. Comput. Linguist. **24**(3), 470–500 (1998)
28. Salton, G., McGill, M.: Introduction to Modern Information Retrieval. McGraw-Hill, Inc., New York (1986)
29. Witten, I., Frank, E., Hall, M.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, San Francisco (2011)
30. Wu, M., Fuller, M., Wilkinson, R.: Using clustering and classification approaches in interactive retrieval. Inf. Process. Manage. **37**(3), 459–484 (2001)
31. Yatani, K., Novati, M., Trusty, A., Truong, K.N.: Review spotlight: a user interface for summarizing user-generated reviews using adjective-noun word pairs. In: Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems, CHI 2011, pp. 1541–1550 (2011)

# A Comparison of Five HSV Color Selection Interfaces for Mobile Painting Search

Min Zhang<sup>1()</sup>, Guoping Qiu<sup>1</sup>, Natasha Alechina<sup>1</sup>, and Sarah Atkinson<sup>2</sup>

<sup>1</sup> School of Computer Science, University of Nottingham, Nottingham, UK  
`{mzz, qiu, nza}@cs.nott.ac.uk`

<sup>2</sup> Human Factors Research Group, University of Nottingham, Nottingham, UK  
`sarah.atkinson@nottingham.ac.uk`

**Abstract.** Color selection is a common task in a plethora of mobile applications. Although a variety of color palettes emerge, there are no design guidelines or studies of the use of color palette interfaces for drawing on touch-screen phone. We are particularly interested in drawing queries to search for paintings on mobile phone. In this paper, we classified the color palette interfaces into several categories. We report results of a systematic experiment with 41 participants using five different types of HSV color palettes for the task of drawing a painting to search on a mobile phone. We investigate which color palette(s) enable users to complete task faster, how good these resulting drawings were for searching, and what were user experiences. Users' drawing behavior is also discussed.

**Keywords:** Color palette interface · User study · Painting search · User interface design · Drawing-to-search · Mobile application

## 1 Introduction

Color is a salient visual attribute for both paintings [1] and our daily life. When searching paintings on the keyword-based image search engine (e.g. Google), we found that most of queries describe the expected paintings by use of color names such as '*golden lady in green*', '*yellowish hat*' etc. However, the huge '*semantic gap*' [4] makes search results unsatisfactory. Moreover, it is difficult for users to describe some paintings and pictures in several keywords, such as abstract paintings.

With the proliferation of the touch-screen mobile devices, people are able to express their intention more freely. Zhang et al. [6] proposed that most users could draw out rough color-sketches of previously-seen paintings from their memory. As one of the image retrieval paradigms, Banfi [5] concluded that *Query-by-Drawing* (QbD) was the most flexible way for users to present their search ideas. There are many studies on QbD algorithms [7, 8] which focused on the search accuracy and efficiency, however, little is known about the design guidelines of the natural and intuitive user interfaces that facilitate user's drawing.

For the small-screen mobile phone, drawing color patches [9] roughly is more reasonable [10]. One of the most important tasks for drawing is color selection. There are a variety of ways to select colors, and the most common method is to directly

manipulate a color palette [3]. The color palette usually includes a color model and an interface to represent colors to users. Several studies were conducted to explore the effects of different color models on color matching task. Schwarz et al. [11] compared five color models: RGB, HSV, LAB, YIQ and Opponent, and they found that RGB color model was the fastest to use yet least accurate, while HSV color model was the most accurate to use. However, these arguments were only concluded from conducting the color matching experiment, which ignored the color selection interface for drawing purpose.

This work aims to focus on the user interface of color palette, by focusing on a specific color model and a current *Query-by-Drawing* algorithm, to explore the influence of different color palette interfaces on color selection for drawing query on mobile phone. Specifically, an experiment was conducted to collect task performance and users' attitudes on five different color palette interfaces.

## 2 Related Color Palettes Study

The color palette has been widely explored by the scientific visualization researchers [12], with aims of displaying information efficiently by combining proper colors. Looking through current mobile application (referred to ‘App’ in the remainder of this paper) market and computer software, including mobile drawing Apps, coloring games for kids, note-taking Apps, and photo editing software and Apps, we found that a plethora of different color palettes were adopted by different developers according to their individual preference. To the best of our knowledge, there is no design guideline about making use of color palettes for drawing on touch-screen mobile devices, especially for image search by drawing mobile App.

As mentioned earlier, some previous work has investigated the effects of color model on the speed and accuracy of color selection in color matching experiment [11, 13, 14, 18], in which participants choose a closest color to a given color shown on the computer screen by use of the color palettes with different color models. Besides Schwarz et al.’s study [11], Everly and Mason [14] also assessed the usability and accuracy of four color selectors via exploring participants’ performance on color matching. Dogulas and Kirkpatrick [18] found that the color space had little impact on color selection. However, the color matching task is out of context of the real-world drawing scenario in which choosing one color might be influenced by the surrounding or background colors.

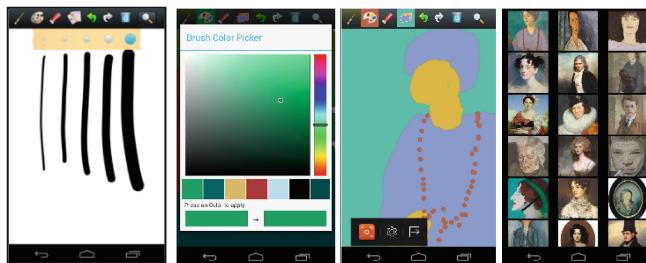
There are a few studies on the color palettes used in Content-based Image Retrieval (CBIR). For example, Broek et al. [15, 16] reviewed ten color selection interfaces for CBIR and categorized them into three groups: *slider-bars*, *discrete* color matrix, and *continuous* color square. However, their work was conducted on the computer-based CBIR systems almost 20 years ago. Their limitations include: (1) non-intuitive interaction of color selection by using mouse; (2) out-of-date color palette interfaces; (3) not representative of the current mobile device capability. Zhang et al. [17] compared two discrete color palettes with two continuous color palettes (HSV square and HSV color wheel) on mobile phone and concluded that

participants preferred the color palette with full-choice of colors during drawing search query. Inspired by Zhang et al.'s work [17], our purpose is to explore which kind(s) of continuous color palette interface work best for a *Drawing-to-Search* mobile App. We also adopt HSV color model in this work, because it is perceptually intuitive and is used by lots of current mobile applications.

Current color research all define a color in a triplet of three parameters, hence selecting a color means finding a location in a three-dimensional (3D) space. We divide the continuous color palette interfaces into '*1D-1D-1D*' and '*1D-2D*'. In '*1D-1D-1D*' interface, each one of three parameters is controlled by one motion (such as a trio of 1D sliders); while in '*1D-2D*' interface, one motion simultaneously change two parameters and one for another parameter, e.g. color palettes shown in Fig. 1. In this paper, we focus on comparing the '*1D-2D*' color palette interfaces based on the human-oriented HSV (*Hue-Saturation-Value*) color model, because this pair of interface and color model is frequently used in both current color selection tools and the graphics literature.

### 3 Query-by-Drawing Application

We built five Client/Server-based drawing Apps on Android phone, and each App provided the same basic drawing tools except the color picker. The basic drawing tools include brushes with five different sizes (10, 20, 30, 40 and 60 pixels), '*Brush color picker*', erasers with five different sizes (10, 30, 40, 50 and 70 pixels), '*Canvas color picker*', '*Undo*', '*Redo*', and '*Trash*' on the top of canvas (see Fig. 1, left). We separated brush color and canvas color to facilitate user to create and refine drawings easily, and the same color selection interface is used for '*Brush color picker*' and '*Canvas color picker*' of each App. Each color palette interface could keep 20 recently used colors.



**Fig. 1.** The screenshots of drawing and result-viewing interface of our *Query-by-Drawing* App.

When users complete one drawing, they tap the '*Search*' button to submit their drawings to our server, which will extract drawing's color features and match them to that of database images by using *Fast Multi-resolution Image Querying* algorithm [2]. Top fifty-one matched images, three image thumbnails on a line and 17 lines in total, will be returned from server to user's mobile phone and displayed in a descending order

of visual similarity. That is, the most similar painting is at the top-left (e.g. Fig. 1, right). The reason behind showing fifty-one results is that users usually view results on the first few pages and fifty-one image thumbnails are on the first three screens. The database contained 400 portrait paintings downloaded from *BBC Your Paintings* website<sup>1</sup>.

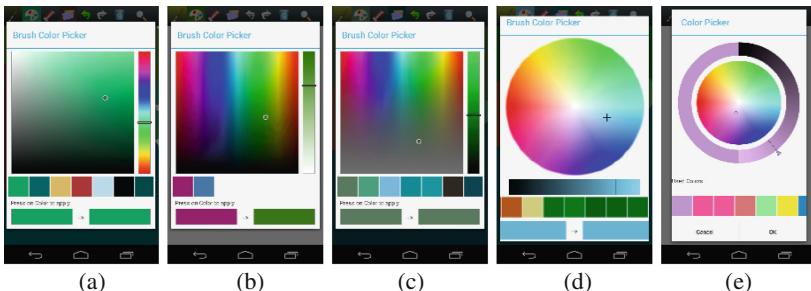
## 4 Experiment

The ‘1D-2D’ color palette interfaces could be designed in a variety of ways in terms of the combinations of three axes of HSV color model and the shapes to display these combinations. We are interested in (1) if the combinations of three parameters of color model have measurable effects on users’ color selection; and we also want to see (2) how visual representations of parameter(s) influence users’ performance on drawing.

### 4.1 Experimental Design

Based on the results of reviewing 140 drawing Apps, 22 coloring Apps, and 18 photo-editing Apps available on Google Play store, we adopted four commonly used shapes of color palette interface, including *square*, *wheel*, *ring* and *bar*. The *square* and *wheel* were used for displaying the ‘2D’ combinations, and *ring* and *bar* were for the ‘1D’.

The *bar-square* interface is explored to answer the first research question. For ‘1D-2D’ interfaces, there are three combinations of HSV color model: H-SV, S-HV, and V-HS. The ‘2D’ is displayed as a *square*, and ‘1D’ is controlled by a vertical slider *bar*. These three combinations were named as *H-SVsquare*, *S-HVsquare*, and *V-HSsquare*, as shown in Fig. 2, and the comparison of these three interfaces was **comparison ‘C1’**.



**Fig. 2.** Five color palette interfaces: (a) *H-SVsquare*, (b) *S-HVsquare*, (c) *V-HSsquare*, (d) *V-HSwheel*, (e) *Vring-HSwheel*

As for the second research question, firstly, we built a slider *bar* and a *ring* of *V* (*Value*) with HS color *wheel*, which were formed into *V-HSwheel* and *Vring-HSwheel* as shown in Fig. 2. We defined **comparison ‘C2’** as *V-HSwheel* vs. *Vring-HSwheel*.

<sup>1</sup> <http://www.bbc.co.uk/arts/yourpaintings/>.

Secondly, we were also interested in looking at two different representations of HS with slider bar: *square* vs. *wheel*, that is, **comparison ‘C3’: V-HS*square* vs. V-HS*wheel***.

The experiment adopted a within-subjects design. The independent variable is ‘*Interface Type*’ (five Apps), and the dependent variables include performance measures and self-rated satisfaction and interface usability obtained from questionnaires.

**Participants.** Forty-one college students (mean age 19 years, s.d. 1.88) have volunteered to take part in the experiment. All participants reported having experience of using touchscreen mobile phone and none of them self-indicated as drawing-expert or artist. All participants were novice users of our Apps and right-handed.

**Material and Stimuli.** Two Google Nexus 4 (133.9 mm × 68.7 mm) mobile phones with Android 4.4 operating system were used: One mobile phone was used to present the reference painting to participants; and five Apps and *SCR Screen Recorder Pro* were installed on the other phone on which participants would draw. Both phones were adjusted to the same level of brightness to reduce the device difference effects.

The choice of painting as stimulus followed the following rationales: (1) The painting is not well-known; (2) The painting has at least six rough color blocks, which contains the color widely used in literature and agreed as the basic colors [18], e.g., blue, red, and green etc. We believe the painting *Old man* (shown in Fig. 3) is sufficient to evaluate user’s color selection performance with at least seven gist color blocks.

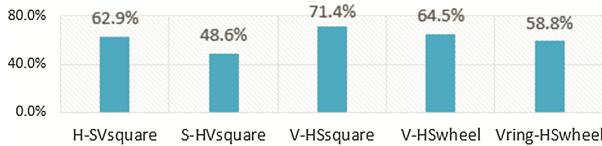


**Fig. 3.** The reference painting *Old man* used for copying task.

## 4.2 Procedure

Only one participant involved in study each time. Participant was briefly informed of the aim and procedure of our study, then he/she was instructed to complete the pre-task questionnaire before a tutorial on how to use our mobile Apps. After a practice session (for eliminating individual difference effects), each participant was asked to complete the copying task (copying painting *Old man* shown in Fig. 4) by use of five Apps in a random order to minimize the practice effects. One drawing was required on each App. We encouraged participants to draw the rough color sketches with accurately tracing color and proportion of the original painting. At the end of each drawing session, participant was asked to rate on the usability with 7-point Likert scale in terms of the drawing speed, the level of ease-of-use, the similarity of his/her drawing compared to the reference painting, and the overall preference on the color selection interface.

After completing all tasks participants were interviewed to specify which interface they believed to be their favorite one among five Apps, and also the preferred



**Fig. 4.** The Percentage of drawings that get the *Old man* on top 51 result lists.

*square*-based color palette and *wheel*-based color palette. The time spent on drawing was captured by *SCR Screen Recorder Pro*. And the position of *Old man* in search results for each query was recorded by the experimenter. All procedures were video-recorded.

## 5 Results and Discussions

206 drawn-queries were produced with relevant retrieval results and drawing-videos recorded. The time taken to complete a drawing, the retrieval result, and the self-rated satisfaction with both drawing process and color selection interfaces were analyzed.

Participants spent the shortest time on ***V-HSwheel*** (mean = 4 m 31 s) to complete one drawing, followed by ***Vring-HSwheel*** (mean = 4 m 47 s), and the longest time taken on ***V-HSsquare*** (mean = 5 m 46 s). A paired samples t-test revealed that the difference between ***V-HSwheel*** and ***V-HSsquare*** was highly significant ( $t(40) = -2.786, p = .008$ ), whereas no significant difference was found between ***V-HSwheel*** and ***Vring-HSwheel*** ( $t(40) = -1.089, p = .283$ ). A repeated measures ANOVA with a Greenhouse-Geisser correction indicated that there were no significant differences ( $F(1.620, 64.816) = 0.197, p = .775$ ) among the time taken to complete one drawing on ***H-SVsquare***, ***S-HVsquare*** and ***V-HSsquare***. We also found that the time spent on drawing a query gradually decreased as participants' experience gained.

Figure 4 illustrates the proportion of drawings on each App that got the *Old man* in top 51 results. The ***V-HSsquare*** worked best (71.4 % drawings succeeded) among five Apps, and ***V-HSwheel*** (64.5 %) was better than ***Vring-HSwheel*** (58.8 %).

Regarding to the usability of color palette interfaces and the user preference, user's self-rated Likert-type scales were collected from questionnaire, were statistically analyzed by Friedman tests and Wilcoxon tests with a significant level of 0.01. The Friedman tests were carried out for the **comparison C1**. There were no statistically significant differences in the interface usability (first seven feature rows illustrated in Table 1) and the preference among three HSV axes combination interfaces (***H-SVsquare*** vs. ***S-HVsquare*** vs. ***V-HSsquare***), as shown by column **C1** in Table 1.

For participants' attitudes to drawing process and preference, the Wilcoxon's singed rank tests indicated that ***V-HSwheel*** was highly significantly better than ***Vring-HSwheel*** in terms of participants' satisfaction rankings on *Relaxing-to-draw* ( $Z = -3.123, p = .002$ ), *Quick-to-draw* ( $Z = -2.701, p = .007$ ), *Easy-to-draw* ( $Z = -3.257, p = .001$ ), *similarity of drawing* ( $Z = -2.853, p = .004$ ), and *easy-to-pick wanted color* ( $Z = -2.755, p = .006$ ). While comparing ***V-HSwheel*** and ***V-HSsquare***, a highly significant difference ( $Z = -2.644, p = .008$ ) was only found

**Table 1.** The statistics analysis of self-rated usability, user preference on color palette interface.

Features\Comparison	C1	C2	C3
Relaxing-to-draw	$\chi^2(2) = 2.048, p = .359$	$Z = -3.123, \mathbf{p = .002}$	$Z = -2.263, p = .024$
Quick-to-draw	$\chi^2(2) = 0.290, p = .865$	$Z = -2.701, \mathbf{p = .007}$	$Z = -1.995, p = .046$
Easy-to-draw	$\chi^2(2) = 0.587, p = .746$	$Z = -3.257, \mathbf{p = .001}$	$Z = -1.498, p = .134$
Level of similarity of query to the origin painting	$\chi^2(2) = 1.355, p = .508$	$Z = -2.853, \mathbf{p = .004}$	$Z = -0.190, p = .849$
Enough color choices	$\chi^2(2) = 5.291, p = .086$	$Z = -1.034, p = .301$	$Z = -2.644, \mathbf{p = .008}$
Easy-to-pick the wanted color	$\chi^2(2) = 2.947, p = .229$	$Z = -2.755, \mathbf{p = .006}$	$Z = -2.274, p = .023$
Quick-to-get the wanted color	$\chi^2(2) = 0.938, p = .625$	$Z = -1.523, p = .128$	$Z = -2.543, p = .011$
Color palette preference	$\chi^2(2) = 5.291, p = .071$	$Z = -0.485, p = .628$	$Z = -1.812, p = .070$

in the ratings of *enough color-choices*, that is, participants thought that **V-HS<sup>wheel</sup>** provided more color choices than **V-HS<sup>square</sup>**. But in fact, both interfaces supply the same number of color choices.

There were no statistically significant differences in participants' ratings on interface preference. However, from answers to question about which interface was preferred in each group of *bar-square* and *wheel-based* color palettes, most people chose **H-SV<sup>square</sup>** as their favorite *bar-square* color picker, and **Vring-HS<sup>wheel</sup>** as their favorite *wheel-based* color palette. And they preferred **H-SV<sup>square</sup>** color interface overall.

It is found that each drawing contains 6 ~ 8 colors on average. A paired samples t-test revealed that the difference between the number of colors used for drawing on **V-HS<sup>wheel</sup>** (mean 6.87) and **V-HS<sup>square</sup>** (mean 7.46) was highly significant ( $t(40) = -3.438, p = .001$ ).

It is observed that how participants draw varied significantly, however, several common issues of drawing behavior emerged. First, participants with high-level drawing skills, e.g. sketching or oil painting, focused on small details of painting during the first drawing session. For example, *Participant #16* and *#20* drew both outline of the coat and face details e.g. eyes, eyeglasses, and mouth etc., as shown in Fig. 5. Second, most of users followed the same drawing style as that of using pencil-and-paper, that is, (a) sketching an outline firstly and filling colors afterwards, e.g. the screenshot of the first sketching process of *Participant #32* as shown in Fig. 5, right; (b) Rubbing out the mistakes and then drawing the correct one rather than drawing over the mistakes directly.



**Fig. 5.** First drawings from *Participant #16, #20, #32*.

## 6 Conclusions

In this paper, we proposed a method of categorizing the continuous color palette interfaces into two classes: ‘*1D-1D-1D*’ interface and ‘*1D-2D*’. By focusing on ‘*1D-2D*’ interface with HSV color model, an experiment (with 41 participants) was designed and carried out to explore the influence of both combinations of the three axes of HSV color model and shape representing these combinations on users’ query-drawing performance. Specifically, we compared three groups of color selection interfaces: three *bar-square* color palettes, *bar-wheel* vs. *ring-wheel*, and *bar-wheel* vs. *bar-square*. The experimental results suggest that shape has more effects on users’ performance than color model axes combinations. In general, *wheel*-based interface leads to faster drawing than the *square*-based one; and the usability rating on *ring-wheel* is higher than that on *bar-wheel*. Most of participants preferred the H-SV *square-bar* colour selection interface.

One limitation of our study is conducting the copying-task only. However, the fact that our reference painting contains 7 different color blocks makes this preliminary study reliable. We will do more studies on other types of tasks, e.g. landscape drawing.

## References

1. Davidoff, J.B.: *Cognition Through Color*. MIT Press, Cambridge (1991)
2. Jacobs, C.E., Finkelstein, A., Salesin, D.H.: Fast multiresolution image querying. In: *Proceedings of SIGGRAPH 1995* (1995)
3. Douglas, S., Kirkpatrick, T.: Do color models really make a difference? In: *Proceedings of CHI 1996 Conference*, pp. 399–405. ACM Press, New York (1996)
4. Hare, J.S., Lewis, P.H., Enser, P.G.B., Sandom, C.J.: Mind the gap: another look at the problem of the semantic gap in image retrieval. In: *Proceedings of Multimedia Content Analysis, Management and Retrieval* (2006)
5. Banfi, F.: Content-based image retrieval using hand-drawn sketches and local features – a study on visual dissimilarities, Ph.D. Thesis, University of Friburg, Germany (2000)
6. Zhang, M., Atkinson, S., Qiu, G.P., Alechina, N.: Can people finger-draw color-sketches from memory for painting search on mobile phone? In: *MOMM 2014*, pp. 115–118 (2014)
7. Kato, T., et. al.: A sketch retrieval method for full color image database-query by visual example. In: *Proceedings of ICPR 2008*, pp. 530–533 (1992)
8. Springmann, M., Ispas, A., Schuldt, H., Norrie, M.C., Signer, B.: Towards query by sketch. In: *Second DELOS Conference on Digital Libraries* (2007)

9. Giangreco, I., Springmann, M., Kabary, I., Schuldt, H.: A user interface for query-by-sketch based image retrieval with color sketches. In: Baeza-Yates, R., de Vries, A.P., Zaragoza, H., Cambazoglu, B.B., Murdock, V., Lempel, R., Silvestri, F. (eds.) ECIR 2012. LNCS, vol. 7224, pp. 571–572. Springer, Heidelberg (2012)
10. Ivanova, K., Stanchev, P., Dimitrov, B.: Analysis of the distributions of color characteristics in art painting images. *Serdica J. Comput.* **2**(2), 111–136 (2008)
11. Schwarz, M., Cowan, W., Beatty, J.: An experimental comparison of RGB, YIQ, LAB, HSV and opponent color models. *ACM Graph.* **6**(2), 123–158 (1987)
12. Rogowitz, B., Treinish, L.: Why should engineers and scientists be worried about color? *Lloydia Cincinnati* (2009)
13. Henry, P.M., Westland, S., Cheung, T.L.V.: An intuitive color-selection tool. In: Proceedings of 14th Color Imaging Conference (2006)
14. Everly, D., Mason, J.S.: Color selection methods using the color picker (1999). <http://www.otal.umd.edu/SHORE99/jsmason>
15. van den Broek, E.L., Vuurpijl, L.G., Kisters, P., von Schmid, J.C.M.: Content-based image retrieval: color-selection exploited. In: Proceedings of DIR-2002, pp. 38–47 (2002)
16. van den Broek, E.L., Kisters, P.M.F., Vuurpijl, L.G.: Design guidelines for a content-based image retrieval color-selection interface. In: Dutch HCI 2004 (2004)
17. Zhang, M., Qiu, G.P., Alechina, N., Atkinson, S.: User study on color palettes for drawing query on touchscreen phone. In: Proceedings of MHCI 2013 (2013)
18. Douglas, S., Kirkpatrick, A.E.: Model and representation: the effect of visual feedback on human performance in a color picker interface. *ACM Trans. Graph.* **18**(2), 96–127 (1999)

# Computer-Related Attribution Styles: Typology and Data Collection Methods

Adelka Niels<sup>(✉)</sup> and Monique Janneck

Luebeck University of Applied Sciences, Luebeck, Germany  
[{adelka.niels,monique.janneck}@fh-luebeck.de](mailto:{adelka.niels,monique.janneck}@fh-luebeck.de)

**Abstract.** Attribution, i.e. the systematic ascription of causes to effects in situations of failure or success, has so far received little attention in HCI research. Based on a preliminary typology developed in pilot work, we conducted four empirical studies with a total of  $N = 146$  participants using different methods for data collection, including laboratory studies, a mobile diary study, and an online survey. Results show that several typical styles of attributing computer-related failure or success could be identified. Therefore, we propose a typology of six main attribution styles, which are depicted as personas to make them applicable for HCI practice. Methodical issues in computer-related attribution research and implication for research and practice are discussed.

**Keywords:** Attribution · Computer-related attitudes · Computer mastery · Computer failure · User types · Personas

## 1 Introduction: Attribution Research

Causal Attribution research deals with the explanations people find in situations of success and failure for *why* things happened the way they did, and the extent of *control* that people feel they have over external events [1]. The way people explain success or failure can be classified in four dimensions: *locus*, *stability*, *controllability*, and *globality* [e.g. 2, 3].

- *Locus* (internal vs. external) describes whether a person sees internal (“I did not study enough”) or external (“the exam was too difficult/the examiner was unfair”) causes of an event.
- *Stability* (temporally unstable vs. temporally stable) captures whether causes change over time (“This time I failed”) or not (“I always fail”).
- *Controllability* (high control perception vs. low control perception) distinguishes controllable causes (“I could have studied more”) from causes perceived as uncontrollable (“Studying more would not have helped”).
- *Globality* (specific vs. global) describes whether the cause relates to a specific situation (“I just don’t like this subject”) or if it is a global cause (“I never do well in written exams”).

Attribution processes are highly relevant for people’s behavior, emotions, and motivation [1, 4]. For example, attributing a situation of failure within the internal/stable dimensions can lead to shame or humiliation because causes are attributed to the self and seen as

something that cannot be changed. Contrary, internal/instable attributions might also cause self-doubts and self-reproach, but the situation is seen as a singular event that will not necessarily occur again. If a situation of failure is attributed within the external/stable dimension there is less motivation to change. Recurring attributional patterns in different situations and contexts are called *attribution styles*. Attribution styles are considered as part of one's *self-concept*, which represents all of a person's self-referred attitudes [5]. Therefore, attribution styles can be seen as rather stable over time.

We believe that different attribution styles have different influences on user experience and behavior. For example, users with different attribution styles might have quite different explanations for events like system failures, triggering different user responses. Thus, having favorable or unfavorable attribution styles, respectively, might account for differences regarding computer mastery, computer anxiety, or simply different styles of using computers, as has been shown in different studies (for an overview see [6]). Therefore, a detailed knowledge of computer-related attribution styles might help to understand user behavior and difficulties when using computers better.

Even though there is evidence that attribution styles are domain-specific [cf. 6], so far no extensive model of specifically computer-related attributions has been developed. Thus, our research aimed at exploring distinct *computer-related attribution styles*.

A first typology of nine specific attribution styles was identified in two pilot studies, namely a diary study and an online survey [7]. Stereotypical names and exemplary statements were used to illustrate the kind of attitude and behavior that might be associated with the respective attribution style (Table 1).

**Table 1.** Typology of computer-related attribution styles [7]

	<i>Style – Description</i>	<i>Diary</i>	<i>Survey</i>
Success	Realistic – “Sometimes I am successful, sometimes not”	x	x
	Humble – “This time I was lucky”	x	x
	Lucky guy – “Everything I do turns out right”	x	x
	Confident – “I am competent and responsible for my success”	x	x
	The Boss – “Success depends on the system, but I control it”		x
Failure	Realistic – “This time I failed, but don't worry about it”		x
	Shrugging – “Every failure is unique”	x	x
	Confident – “I know it was my fault, but next time I will do better”	x	x
	Resigned – “I never understand what computers do”	x	x

In this paper, the primary objective was to reproduce and refine this typology and therefore present a validated concept, which can be used in further HCI theory and practice. Furthermore, we aimed to explore and compare different data collection methods.

## 2 Research Questions and Methods

In this paper, we investigated the following main research questions and objectives:

- Identification of a main set of computer-related attribution styles that can be reproduced in a variety of different settings, using different methods for data collection;

- Exploration of the differences, advantages and drawbacks of different research instruments and recommendations for further research.

To that end, we investigated computer-related attribution styles by means of four different data collection methods: A *laboratory study* with standardized use situations; a *mobile diary application* to measure attributions in everyday use situations; an *attribution questionnaire* which was filled out by participants as part of a *usability test*; and an *online survey* recording retrospective memories of computer-related situations of success and failure. Participation in all studies was anonymous. Therefore, it is possible that some persons have participated in more than one study. The studies are described in detail in Sects. 3, 4, 5 and 6.

In all four studies the same measure was used in order to compare the results, which had been developed and tested in the pilot studies [7]. In addition to four questions measuring the attributional dimensions of locus, stability, controllability and globality (Fig. 1, based on the Sport Attributional Style Scale, SASS, [8]), participants were asked to briefly describe the cause of failure or success and also rate its significance and task difficulty. Furthermore, socio-demographic data (e.g. age, gender, education, general computer use and experiences) were collected.

In this measure, low values for *locus* mean that a person attributes reasons for success (e.g. “I am competent”) or failure (e.g. “It was my fault”) internally, while high values indicate that they attribute reasons to external factors (e.g. “The system is stable and runs well” vs. “The system is to blame”). Low *stability* values mean that causes are believed to change over time (e.g. “This time I was lucky” vs. “This time I failed”), high stability values indicate recurring events (e.g. “I am always right” vs. “I always fail”). Due to the wording of the questionnaire, low values of *controllability* indicate high perception of control (e.g. “success is due to my diligence” vs. “I did not try hard enough”), while high values of controllability indicate low perception of control (e.g. “I was lucky” vs. “I cannot change the situation anyway”). Finally, low values of *globality* indicate that attributions are not generalized to other situations (e.g. “I can handle this specific application well” vs. “I just don’t understand this specific application”), while high values of globality indicate that similar attributions take place in different contexts (e.g. “I always do well with computers” vs. “I never master computer applications”).

<b>1. I would locate the cause of the breakdown ...</b>
internally (I am to blame)    1 2 3 4 5 6 7    externally (the system is to blame)
<b>2. The cause of this breakdown is...</b>
a singular event    1 2 3 4 5 6 7    recurring
<b>3. The cause of the breakdown is...</b>
controllable    1 2 3 4 5 6 7    uncontrollable
<b>4. The cause of this breakdown is likely to promote other breakdowns...</b>
just in this situation    1 2 3 4 5 6 7    in other situations as well

**Fig. 1.** Part of the standardized attribution questionnaire for failure situations [7]

The data was analyzed using hierarchical cluster analyses as an exploratory instrument for discovering structures in raw data [9, 10]. Firstly, we measured each subject's level of attribution per dimension collected over each situation. Secondly, we built a matrix, containing the distance (calculated via Euclidian measures) between the subjects regarding each dimension. After that we clustered each subject or group of subjects together, while keeping the inner cluster variance low, using Ward's method for computing the cluster linkage criterion. Finally, to rule out which cluster solution stands out, we considered the variance changes and the plotted structure (dendrogram) for each data set [11].

### 3 Study 1 - Laboratory Study

#### 3.1 Research Methodology

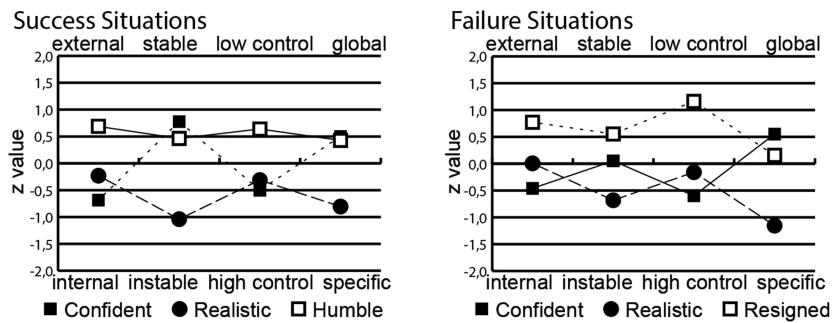
In the first study we investigated computer-related attribution styles in a laboratory setting, conducting usability tests. The participants were asked to edit three task pairings on three different applications or devices, whereby one task of each pairing was easy to solve (situation of success – e.g.: Search for the district office opening hours on a municipal home page) and one task was hard or even insolvable (situation of failure – e.g.: Search for the building regulations of a certain district, which did not exist on the homepage). After each task the participants filled out the attribution questionnaire described in Sect. 2.

#### 3.2 Results

In all, 58 persons participated in the study (48 % female, 52 % male). Mean age was 34 years (range: 17–75 years). The general level of education was quite high (76 % with high school or university degree). On average they had 13 years (range: 0–30 years) of experience in private computer use and 10 years (range: 0–32 years) experience using computers at school or in the workplace. Participants self-rated their computer skills on Likert scales ranging from 1 (low) to 7 (expert) in the various categories: Operating Systems ( $M = 4.34$ ), Internet ( $M = 5.14$ ) and Applications ( $M = 3.82$ ).

We recorded a total of 340 situations, 177 success situations and 163 failure situations. (This imbalance was due to individual perceptions of the outcome of the task: For example, some participants also succeeded in the hard task condition, while others were not successful in the easy task condition).

**Success.** The attributional dimensions are only moderately inter-correlated, thus supporting the construct validity of the research instrument. Merely stability and globality correlate at  $r = 0.51$ . However, this is theoretically plausible: If people believe that success will persist over time they normally also believe that similar situations take place in different contexts.



**Fig. 2.** Clusters for success and failure situations

For success situations, cluster analysis identified the three clusters “*Realistic*”, “*Humble*” and “*Confident*” (Fig. 2).

Persons from Cluster A (“*Realistic*” – “Sometimes I am successful, sometimes not”) attribute reasons for success rather temporally unstable and situation-related, whereby Persons from Cluster B (“*Humble*” – “This time I was lucky”) attribute success to external factors and experience only low levels of control when using computers. Compared to Cluster A and B Persons from Cluster C (“*Confident*” – “I am competent and responsible for my own success”) attribute success temporally stable, globally and internally.

Table 2 shows the mean values for the clusters. ANOVAs were calculated showing significant differences between clusters. Effect sizes (according to Cohen’s classification of  $\eta^2$ , [12]) are high.

**Table 2.** ANOVA results for success clusters

Cluster	A n = 21	B n = 22	C n = 15	F value	p	$\eta^2$
Locus	3.43	4.77	2.77	13.267	<0.000***	0.325
Stability	3.78	5.88	6.32	49.032	<0.000***	0.641
Controllability	1.92	3.01	1.70	9.654	<0.000***	0.260
Globality	3.77	5.34	5.43	16.573	<0.000***	0.376

**Failure.** Regarding inter-correlations, locus and stability correlate at  $r = 0.46$ . Furthermore, locus and controllability correlate low at  $r = 0.28$ . However, this is theoretically plausible: If people see internal causes for a situation they normally also experience higher controllability.

For failure situations, cluster analysis identified the three clusters “*Realistic*”, “*Confident*” and “*Resigned*” (Fig. 2).

Persons from Cluster D (“*Realistic*” – “This time I failed, but I don’t worry about it”) see internal as well as external reasons for failures and believe that these change

over time and depend on a specific situation. Persons from Cluster E (“*Confident*” – “I know it was my fault, but next time I will do better”) have the highest internality values and feel responsible for their failures, but also feel in control of the situation. Persons from Cluster F (“*Resigned*” – “I never understand what computers do”) see external and temporally stable reasons for their failure and feel they have little control over the situation. This attribution style is the most unfavorable of all three clusters and can be compared to the so-called pattern of “learned helplessness” that is observed in patients suffering from depression [cf. 2].

As for success situations, differences between clusters were significant. Effect sizes are high (Table 3).

**Table 3.** ANOVA results for failure clusters

Cluster	D n = 15	E n = 27	F n = 16	F value	p	$\eta^2$
Locus	4.44	3.66	5.73	10.029	<0.000***	0.267
Stability	5.00	5.76	6.28	7.293	0.002**	0.210
Controllability	3.31	2.57	5.55	34.349	<0.000***	0.555
Globality	3.09	5.51	4.94	27.598	<0.000***	0.501

## 4 Study 2 – Field Study with Mobile Diary Application

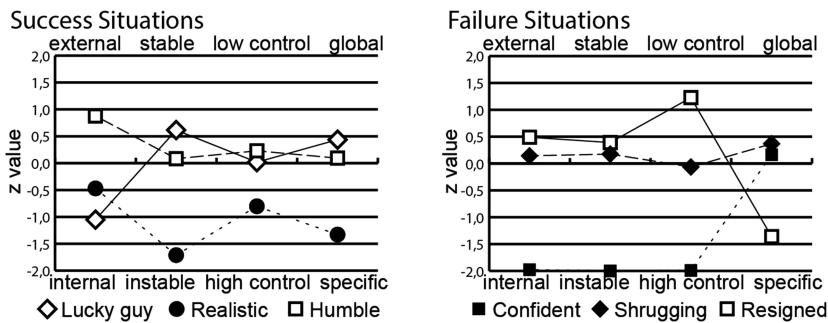
### 4.1 Research Methodology

Participants were asked to record situations of success and failure when using computers in private or workplace situations over a period of 8 weeks by using a mobile diary application. It was up to the participants to decide whether success or failure had taken place. The diary contained ten attributional questionnaire forms each for success and failure [7]. Again, the standard attribution questionnaire was used.

Participants were recruited via extensive publicity measures, including a local newspaper article. However, as diary studies are rather laborious and time-consuming for participants, completion rates are usually low. The survey respondents were able to contact the experimenter at all times. However, we refrained from systematic reminders, which might influence participants’ responses in diary studies [13].

### 4.2 Results

A total of 78 persons participated in the diary study. For data analysis we included all diaries that contained at least one situation of success and failure ( $n = 20$ ). On average participants reported  $M = 3.9$  ( $SD = 3.24$ ; range: 1–10) successes and  $M = 4.0$  ( $SD = 3.35$ ; range: 1–10) failures. 52.6 % of the respondents were female. The mean age of respondents was 34.32 years (range: 19–74 years). The general level of education was quite high (73.7 % with high school or university degree). On average they had 14.95 years (range: 7–21 years) of experience in private computer use and 11.42 years (range: 0–21 years) experience using computers at school or in the workplace.



**Fig. 3.** Clusters for success and failure situations

Participants self-rated their computer skills on Likert scales ranging from 1 (low) to 7 (expert) in the various categories: Operating Systems ( $M = 5.58$ ), Internet ( $M = 6.42$ ) and Applications ( $M = 5.89$ ). Participants reported a total of 159 situations (on average 8 per person). 78 were successes (on average 3.9 per person) and 81 (on average 4.0 per person) failures. 93 of all situations were reported in the workplace and only 66 in private circumstances.

**Success.** All attributional dimensions are only moderately inter-correlated, thus supporting the construct validity of the research instrument.

Just as in study 1, three clusters could be identified for attribution of success (Fig. 3). Clusters G “*Humble*” (with slightly lower values than study 1) and I “*Realistic*” were identical with the clusters in study 1. Cluster H (“*Lucky guy*” – “Everything I do turns out right”) had also been found in previous investigations [7]. Persons in this cluster see reasons for success internally, they feel more in control than persons from cluster A and have the highest values concerning stability and globality, thus displaying a sense of faith that things will simply go right.

Table 4 shows the mean values for the clusters. ANOVAs were calculated showing significant differences between clusters in all dimensions except for controllability. Effect sizes are high, also except for controllability [12].

**Failure.** The attributional dimensions are only moderately inter-correlated (merely controllability and globality correlate at  $r = 0.57$ ).

For failure situations, cluster analysis identified three clusters (Fig. 3). Cluster K (“*Resigned*”) is identical with results from study 1 and already described in Sect. 3. Even though Cluster L (“*Confident*”) looks different at first glance, this cluster represents a typical “*Confident*” attribution style: Persons in this Cluster have the highest

**Table 4.** ANOVA results for success clusters

Cluster	G n = 10	H n = 7	I n = 3	F value	p	$\eta^2$
Locus	4.69	1.55	2.50	47.619	<0.000***	0.849
Stability	5.71	6.48	3.11	13.080	<0.000***	0.606
Controllability	3.63	3.28	1.92	1.263	0.308	0.129
Globality	5.54	6.05	3.42	4.660	0.024*	0.354

internality values and feel responsible for their failure, but also feel in control of the situation. Cluster J (“*Shrugging*” – “Every failure is unique”) had also been found in one of the pilot studies [7]. Persons from Cluster J have medium values in all dimensions; they believe that different situations have unique causes.

Again, ANOVAs were calculated showing significant differences between clusters. Effect sizes are high (Table 5).

**Table 5.** ANOVA results for failure clusters

Cluster	J n = 14	L n = 4	K n = 2	F value	p	$\eta^2$
Locus	5.52	5.96	2.83	7.740	0.004**	0.477
Stability	5.83	6.15	2.67	7.711	0.004**	0.476
Controllability	3.94	5.58	1.50	23.741	<0.000***	0.736
Globality	5.18	2.25	4.83	8.266	0.003**	0.493

## 5 Study 3 – Using Attribution Questionnaires in Usability Tests

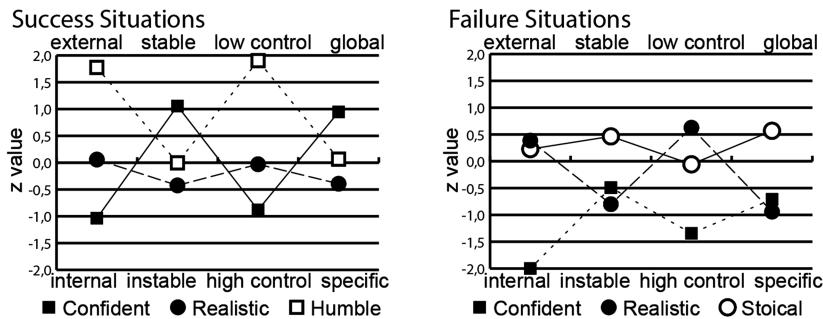
### 5.1 Research Methodology

In this study we investigated computer-related attribution styles as part of different usability tests that were conducted in our laboratory, covering a range of different applications (e.g. online rail-ticket booking system, stock price websites, news websites). After completing the usability tests, participants were asked whether success or failure had occurred during the test and were shown the respective parts of the attribution questionnaire described in Sect. 2. It was up to the participants to decide whether, and—if so—how many successes or failures had taken place in the usability test (up to 3 each could be reported).

### 5.2 Results

In all, 32 persons participated in the study (22 % female, 78 % male). Mean age was 27.42 years (range: 21–64 years). The general level of education was quite high (94 % with high school or university degree). On average they had 16 years (range: 8–22 years) of experience in private computer use and 12 years (range: 4–22 years) experience using computers at school or in the workplace. Participants self-rated their computer skills on Likert scales ranging from 1 (low) to 7 (expert) in the different categories: Operating Systems ( $M = 5.79$ ), Internet ( $M = 6.12$ ) and Applications ( $M = 5.73$ ). Participants reported a total of 107 situations (68 success situations ( $M$  per Person = 1.56; range 0–3) and 39 failure situations ( $M$  per Person = 0.5; range: 0–2)).

**Success.** In this study, we found higher inter-correlations of the attributional dimensions: Locus and controllability correlate at  $r = 0.45$ . Stability and globality correlate at  $r = 0.72$  and controllability and stability correlate at  $r = 0.42$ .



**Fig. 4.** Clusters for success and failure situations

For success situations, cluster analysis identified the three clusters “*Realistic*”, “*Confident*” and “*Humble*” (Fig. 4). Thus, results were identical with the laboratory study described in Sect. 3.

Table 6 shows the mean values for the clusters. ANOVAs were calculated showing significant differences between clusters. Effect sizes are high [12].

**Table 6.** ANOVA results for success clusters

Cluster	M n = 20	N n = 8	O n = 4	F value	p	$\eta^2$
Locus	3.28	1.65	5.83	31.610	<0.000***	0.686
Stability	5.63	6.94	6.00	9.836	0.001**	0.404
Controllability	2.87	1.38	6.25	29.229	<0.000***	0.668
Globality	5.32	6.81	5.83	7.167	0.003**	0.331

**Failure.** Regarding inter-correlations, again stability and globality correlate at  $r = 0.73$  and locus and controllability correlate at  $r = 0.46$ .

For failure situations, cluster analysis identified the previously known clusters of P “*Realistic*” and R “*Confident*” as well as a new cluster we termed “*Stoical*” (Fig. 4). Compared to previous studies, however, persons in Cluster P had lower values in the control dimension. Cluster Q constitutes a new style not previously found in any of the other studies. Persons from Cluster Q (“*Stoical*” – “It’s all the same over and over again”) have high values regarding stability and globality, and medium values regarding locus and controllability. Thus, they perceive persistent causes of failures over time or in different situations while displaying a similar sense of controllability and internality as the “*Realists*”.

Again, ANOVAs were calculated showing highly significant differences between clusters. Effect sizes are high (Table 7).

**Table 7.** ANOVA results for failure clusters

Cluster	P n = 8	Q n = 17	R n = 3	F value	p	$\eta^2$
Locus	6.00	5.71	1.00	25.461	<0.000***	0.671
Stability	4.19	6.15	4.67	6.795	0.004**	0.352
Controllability	5.69	4.29	1.67	5.847	0.008**	0.319
Globality	2.56	5.44	3.00	13.589	<0.000***	0.521

## 6 Study 4 – Retrospective Online Survey

### 6.1 Research Methodology

In this study computer-related attributions were measured by means of a retrospective online questionnaire. Participants were asked to remember their latest computer-related situations of success and failure and fill out the attribution questionnaire while reconsidering these experiences. The call for participation was distributed virally via e-mail and social networking sites.

### 6.2 Results

In all, 90 persons participated in the study. We included only completed questionnaires into data analysis that contained one situation of success and failure ( $n = 30$ , 32 % female and 68 % male). Mean age was 28 years (range: 20–45 years). The general level of education was quite high (80 % with high school or university degree). On average they had 14.4 years (range: 5–21 years) of experience in private computer use and 10 years (range: 0–21 years) experience using computers at school or in the workplace. Participants self-rated their computer skills on Likert scales ranging from 1 (low) to 7 (expert) in the different categories: Operating Systems ( $M = 5.97$ ), Internet ( $M = 6.27$ ) and Applications ( $M = 5.60$ ). Participants reported a total of 60 situations (30 success situations and 30 failure situations).

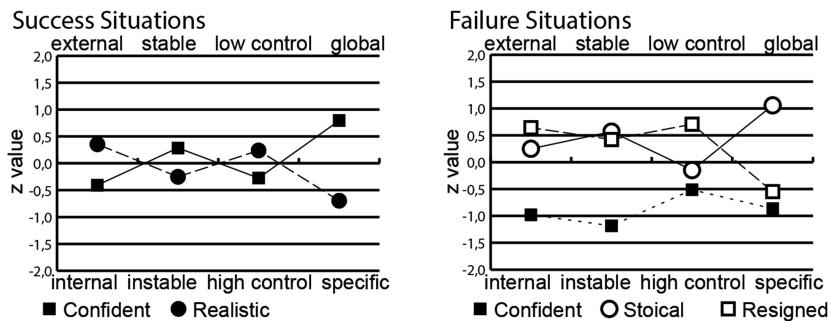
**Success.** As was observed in study 1, only stability and globality correlate at  $r = 0.50$ .

In this study only two clusters appeared for situations of success: “*Realistic*” and “*Confident*” (Fig. 5). Both of them have already been described in the previous sections.

Table 8 shows the mean values for the clusters. ANOVAs were calculated showing significant differences and high effect sizes regarding locus and globality [12]. Differences regarding stability and controllability are not significant.

**Failure.** Regarding inter-correlations, locus and stability correlate at  $r = 0.58$  and stability and globality correlate at  $r = 0.71$ .

In situations of failure three clusters could be identified (Fig. 5). The clusters “*Resigned*” and “*Confident*” were identical with previous studies. Furthermore, the “*Stoical*” style newly identified in study 3 was also present in this study.

**Fig. 5.** Clusters for success and failure situations**Table 8.** ANOVA results for success clusters

Cluster	S n = 16	T n = 14	F value	p	$\eta^2$
Locus	3.81	2.43	4.891	0.035*	0.149
Stability	5.44	6.29	2.252	0.145	0.074
Controllability	2.63	1.79	2.042	0.164	0.068
Globality	3.13	6.50	37.651	<0.000***	0.573

Compared to previous studies, persons from Cluster U (“Resigned” – “I never understand what computers do”) show lower values regarding globality. This was also observed in the diary study and might be due to the method of data collection.

As for success situations, ANOVAs were calculated showing significant differences between clusters. Effect sizes are high (Table 9).

**Table 9.** ANOVA results for failure clusters

Cluster	W n = 9	V n = 12	U n = 9	F value	p	$\eta^2$
Locus	6.33	5.50	2.89	11.094	<0.000***	0.451
Stability	6.33	6.67	2.67	22.466	<0.000***	0.625
Controllability	5.78	3.83	3.00	4.405	0.022*	0.246
Globality	2.44	6.42	1.67	50.767	<0.000***	0.790

## 7 Discussion

### 7.1 Interpretation of Results and Resulting Typology

The goal of this paper was to validate and refine the typology of computer-related attribution styles (cf. Fig. 1) that had been developed in two pilot studies [7] and to investigate whether these styles are reproducible through the use of different methods of data collection. To this end, four elaborate studies with a total of N = 146 participants were conducted (see Table 10).

**Table 10.** Comparison of results of the four Studies.

	Study 1 lab	Study 2 diary	Study 3 usability	Study 4 QN
Total response N	66	78	46	90
Valid response N	58	20	32	30
Female %	46.9	52.6	21.9	32.1
Male %	53.1	47.4	78.1	67.9
Age, M, years	38.2	34.32	27.42	28.37
Range	17–75	19–74	21–64	20–45
Education, M (Scale: 1 no educational degree – 7 university degree)	4.77	6.26	6.12	5.97
Computer experience in years (private)	12.36	14.95	15.67	14.4
Range	0–21	7–21	‘8–22	5–21
Computer experience in years (work)	10.21	11.42	10.97	10.23
Range	0–21	0–21	4–22	0–21
Daily computer use in hours (private)	2.45	2.53	4.33	3.48
Range	0–21	0–6	1–10	0–15
Daily computer use in hours (work)	3.7	4.26	6.12	3.9
Range	0–10	0–9	1–11	0–14
Self-assessed computer skills OS, M (Scale: 1 low – 7 expert)	4.17	5.58	5.79	5.97
Self-assessed computer skills Internet, M (Scale: 1 low – 7 expert)	4.98	6.42	6.12	6.27
Self-assessed computer skills applications, M (Scale: 1 low – 7 expert)	3.66	5.89	5.73	5.6
<b>Total situations</b>	<b>387</b>	<b>159</b>	<b>107</b>	<b>60</b>
Situations of success	197	78	68	30
Situations of failure	190	81	39	30

The four studies comprised a *laboratory study* explicitly evoking situations of success and failure by giving participants solvable and non-solvable tasks (Sect. 3), a *mobile diary study* where participants reported real use situations over a period of several weeks (Sect. 4), a study investigating computer-related attribution styles in various *usability tests* (Sect. 5), and a *retrospective online survey* asking participants to remember their last computer-related situations of success and failure (Sect. 6).

To sum up, results of the four studies confirm the typology of computer-related attribution styles that had been developed in the pilot studies [6]. Even though very different methods of data collection were used, the same attribution styles emerged over and over again, thus supporting the assumption that people indeed display stable, specific computer-related attribution styles.

Overall, a total of ten attribution styles were identified (five of them related to success or failure situations, respectively). While all styles described in the pilot studies could be reproduced except for one (“*The Boss*”, see Fig. 1), a new style, called the

“*Stoical*”, emerged in the usability study (study 3) as well as the retrospective survey (study 4). The refined typology resulting from our studies is shown in Table 11. The numbers indicate how often the specific style appeared in each study.

As can be seen from this overview, especially six styles emerged in almost all studies: In situations of success, these styles include “*Realistic*” (41.29 %), “*Confident*” (27.36 %) and “*Humble*” (24.38 %); in situations of failure, “*Confident*” (25.98 %) “*Resigned*” (23.04 %) and “*Realistic*” (17.16 %) reappear the most frequently. (Also, the “*Shrugging*” style was among the most frequent, however it emerged in only one of the four present studies. Furthermore, the “*Realistic*” style is more expressive regarding the behaviors associated with it).

Among the finally chosen, the “*Confident*” styles can be seen as *favorable* attribution patterns, as these persons experience ample control when working on computer-related tasks and feel confident to handle even difficult situations. On the other hand, the “*Humble*” and “*Resigned*” styles can be seen as *unfavorable* styles: People with these attributional patterns feel that success or failure are due to external factors and there is little they can do to change the situation – a pattern of helplessness. The “*Realistic*” styles are situated in between these extremes: Neither overly confident nor overly pessimistic.

**Table 11.** Refined typology of computer-related attribution styles

		<i>Pilot Studies [6]</i>		<i>Studies 1-4</i>				
	<i>Style</i>	<i>Diary</i>	<i>QN</i>	<i>Lab</i>	<i>Diary</i>	<i>Usability</i>	<i>QN</i>	<i>%</i>
Success	<b>Realistic</b>	3	20	21	3	20	16	41.29
	<b>Humble</b>	2	11	22	10	4		24.38
	Lucky guy	1			7			3.98
	<b>Confident</b>	3	15	15		8	14	27.36
	The Boss		6					2.99
Failure	<b>Realistic</b>		12	15		8		17.16
	Shrugging	7	19		14			19.61
	<b>Confident</b>	1	11	27	2	3	9	25.98
	<b>Resigned</b>	3	15	16	4		9	23.04
	<b>Stoical</b>					17	12	14.22

In our view, these main styles especially deserve attention in further HCI research. To utilize them in design processes, we developed *personas*, which are described in Sect. 7.3.

## 7.2 Methodical Discussion

Another aim of our research was to compare different methods of investigating computer-related attribution styles. As all methods – laboratory tests as well as field studies and online surveys – yielded very similar results we conclude that a wide range

of research methods can be utilized for attribution research. In the following paragraphs the advantages and drawbacks of the different methods and differences in terms of the attribution styles emerged with each method are discussed.

*Laboratory studies* usually require a rather high effort for preparation and conduction of the tests. They also require some time and effort by the participants, who need to show up at a certain time and place. One big advantage, on the other hand, is the high completion rate, as participants usually work under supervision of the experimenter. Standardized use situations create a very similar experience for all participants. Therefore, this was a valuable method to explore computer-related attributions in an early stage of this research. The drawback, however, is that the situations are somewhat artificial and unrelated to the participants' normal use habits, which might result in reduced intensity and significance of the experience. Nonetheless, the laboratory study was the only method that yields no more or no less than the six main attribution styles. Furthermore, an interesting finding is that in this study in situations of failure the "Confident" style (47 %) emerged the most often (compared to the other studies) even though participants had less computer experience, less self-assessed skills and used computers less frequently than the participants in the other studies. Whether this finding is related to the method as such or to other factors cannot be clearly answered.

As an alternative, including attribution questionnaires in *usability studies* also turned out to be a very feasible way of data collection. Short scales measuring the four attribution dimensions like the ones we used could be easily included in usability tests to systematically measure attributions. Of course usability tests might also be conducted outside the laboratory. Concerning the attribution styles in situations of success the "Realistic" (63 %) style appeared most often compared to the other studies. Unlike the laboratory study the gender distribution was not balanced (78 % male). Maybe men show the "Realistic" style more often, as this was also the case in the questionnaire study (53 % Realistic style, 68 % male). Furthermore, a new style—the "Stoical"—emerged in both studies. Again, this might be a style especially shown by younger men.

The main advantage of *diary studies* in general is that the participants record real situations in a prompt and detailed manner [7, 13, 14]. However, participants need to be highly motivated to actually keep their diary for a longer period of time and remind themselves to record relevant situations in a timely manner during their everyday activities, when they have other more important things to do. As a result, dropout rates are usually very high [13, 14].

In our study we used a digital diary application – a web app to be used on desktop and mobile devices, which was specifically tailored to our attribution questionnaire. The goal was to make it easier for participants to record their experiences, e.g. using their smartphones when on the way. However, developing such an application requires a long preparation time, quite some considerable effort and relatively high costs. Furthermore, some interference between the subject of evaluation and the method of data collection occurred: Some participants reported technical problems with the diary application and were, therefore, not motivated to carry on with the diary. In the end, a simple paper-based diary might be a better choice, making it both easier for participants

to record their use experiences and causing less effort for researchers. In any case it is notable that compared to the pilot study, when a paper-based diary had been used, neither participation nor dropout rates were lower than in our present study using a mobile diary application [cf. 7]. In comparison to the other studies, in situations of success the “Confident” style did not emerge in the mobile diary studies. Instead, we found the “Lucky guy” style, which also appeared in the paper diary study. Likewise, in situations of failure both diary studies show the “Shrugging” attribution style. These differences might be due to the fact that in diary studies people report everyday use situations: Possibly, the perception of everyday situations differs from given tasks or situations. It would be interesting to investigate whether the “reality” of a situation influences attribution styles.

Using an *online survey* is a fast, low-cost and easy method to measure attribution styles for both experimenters and participants, giving them the freedom to attend the survey at almost any time and in any location. In the pilot study, a different questionnaire had been used, presenting participants with fictive scenarios of computer-related successes and failures and asking them to envision these situations. Several participants commented afterwards that the situations had appeared artificial to them and they found it difficult to relate to them.

Therefore, in the present online study we took a different approach and asked participants to remember their last computer-related situations of success and failure. While the advantage is that they reported real use situations, a possible drawback is that recalling attributions after some time has elapsed may result in distorted perception [15]. This might also explain that an additional cluster emerged from the data in situations of failure, while only two clusters could be replicated in situations of success. In situations of failure, persons with a “Resigned” attribution style also refer the cause for their failure to a specific situation just like in the diary study, possibly indicating that “real”, spontaneous situations differ from more “artificial”, given tasks. Nevertheless, the results are mainly comparable with the other studies, so overall we hold this to be a feasible method to investigate computer-related attributions.

The completion rate in this study was 30 %. While this is a common value for online studies, it is striking that most participants dropped out at the item asking for a short description of the last computer-related situation of success. We suspect that participants found this too tedious and time-consuming to fill out. We included this item to understand the users’ experiences better and to be able to investigate possible relations between use context and attribution styles [cf. 16]. However, to encourage more participation it might be feasible to do without this item or offer participants a list of more generic descriptions to choose from (e.g. “related to operation system”, “web application” etc.).

In this paper we cannot clearly answer the question whether the differences in attribution styles in different studies are related to the method of data collection or to other factors concerning the sample composition (e.g. age, gender, experience, skills). However, further analyses of the data indicate that socio-demographic factors play a role [16]. It would be interesting to use these methods on one sample group in future investigations to make the differences clear.

### 7.3 Implications and Future Work

The typology of computer-related attribution styles can be used in HCI research and practice to understand better why users think, feel, or behave in a certain way. It is easy to imagine that a “humble” user will behave differently than a “confident” user and thus might react differently to system design [17]. Thus, design principles could be developed to support different types of users in a specific way. Furthermore, including attribution styles as personal traits in usability studies could help to understand and interpret results: E.g., the number of bugs reported could be related to attribution styles, or participants with more unfavorable styles will probably experience more stress during usability studies.

To make the typology of attribution styles more applicable especially for HCI practice, we developed *personas* (Fig. 6 for situations of success and Fig. 7 for situations of failure) to represent the six most central styles (see Sect. 7.1).

Personas are vivid descriptions of fictive yet characteristic user types that can be used in usability engineering processes to envision the future users of an application and keep their needs, attitudes, and prerequisites in mind [18]. Thus, the “attributional personas” are a tool for designers to envisage users with distinct approaches that they are very likely to encounter in later use practice. It should be noted that due to limited space only male personas are pictured here. We are currently developing an extensive set of personas representing both men and women, different age groups, educational backgrounds etc.



**Fig. 6.** Personas for central attribution styles in situations of success.



**Fig. 7.** Personas for central attribution styles in situations of failure.

In our research we are currently investigating the relation of *socio-demographic* variables such as age, gender, computer mastery etc. with attribution styles [16]. Furthermore, we are interested to investigate whether attributional styles influence *system evaluations* in usability tests. To that end, we are conducting further analyses of the usability data collected in study 3. As a long-term perspective, we seek to develop *design principles* to support users with different attribution styles with tailored system responses.

## References

1. Försterling, F.: Attribution: An Introduction to Theories, Research, and Applications. Psychology Press, Hove (2001)
2. Abramson, L.Y., Seligman, M.E.P., Teasdale, J.D.: Learned helplessness in humans: critique and reformulation. *J. Abnorm. Psychol.* **87**, 49–74 (1978)
3. Weiner, B.: Achievement Motivation and Attribution Theory. General Learning Press, Morristown (1974)
4. Kneckt, M.C., Syrjälä, A.M., Knuuttila, M.L.: Locus of control beliefs predicting oral and diabetes health behavior and health status. *Acta Odontol. Scand.* **57**, 127–131 (1999)
5. Marsh, H.W., Byrne, B.M., Shavelson, R.J.: A multifaceted academic self-concept: Its hierarchical structure and its relation to academic achievement. *J. Educ. Psychol.* **80**, 366–380 (1988)

6. Kelley, H., Compeau, D., Higgins, C., Parent, M.: Advancing theory through the conceptualization and development of causal attributions for computer performance histories. *ACM SIGMIS Database* **44**(3), 8–33 (2013)
7. Janneck, M., Guczka, S.R.: The resigned, the confident, and the humble: a typology of computer-related attribution styles. In: Holzinger, A., Ziefle, M., Hitz, M., Debevc, M. (eds.) *SouthCHI 2013. LNCS*, vol. 7946, pp. 373–390. Springer, Heidelberg (2013)
8. Hanrahan, S.J., Grove, J.R., Hattie, J.A.: Development of a questionnaire measure of sportrelated attributional style. *Int. J. Sport Psychol.* **20**(2), 114–134 (1989)
9. Abonyi, J., Feil, B.: *Cluster Analysis for Data Mining and System Identification*. Birkhäuser, Boston (2007)
10. Bacher, J., Pöge, A., Wenzig, K.: *Clusteranalyse: Anwendungsorientierte Einführung in Klassifikationsverfahren*. Oldenbourg Wissenschaftsverlag, Munich (2010)
11. Gillet, N., Vallerand, R.J., Rosnet, E.: Motivational clusters and performance in a real-life setting. *Motiv. Emot.* **33**(1), 49–62 (2009)
12. Cohen, J.: *Statistical Power Analysis for the Behavioral Sciences*, 2nd edn. Erlbaum, Hillsdale (1988)
13. Ohly, S., Sonnentag, S., Niessen, C., Zapf, D.: Diary studies in organizational research: an introduction and some practical recommendations. *J. Pers. Psychol.* **9**, 79–93 (2010)
14. Alaszewski, A.: *Using Diaries for Social Research*. Sage Publications Ltd., London (2006)
15. Reis, H.T., Gable, S.L.: Event-sampling and other methods for studying everyday experience. In: Reis, H.T., Judd, C.M. (eds.) *Handbook of Research Methods in Social and Personality Psychology*, pp. 190–222. Cambridge University Press, New York (2000)
16. Niels, A., Guczka, S.R., Janneck, M.: Computer-related Causal attributions: the role of sociodemographic factors. In: *Proceedings of the 5th AHFE Conference*, 26–30 July 2015 (2015)
17. Phelps, R., Ellis, A.: Overcoming computer anxiety through reflection on attribution. In: Williamson, A., Gunn, C., Young, A., Clear, T. (eds.) *ASCILITE*, pp. 515–524. UNITEC Institute of Technology, Auckland (2002)
18. Cooper, A., Reimann, M.: *About Face 2.0: The Essentials of Interaction Design*. Wiley, New York (2003)

# Reciprocity in Rapid Ethnography

## Giving Back by Making the Small Things Count

Pieter Duysburgh<sup>1()</sup> and Karin Slegers<sup>2</sup>

<sup>1</sup> iMinds – Digital Society, SMIT – VUB, Pleinlaan 9, 1050 Brussels, Belgium  
pieter.duysburgh@vub.ac.be

<sup>2</sup> CUO, KU Leuven/iMinds, Parkstraat 45, bus 3605, 3000 Leuven, Belgium  
karin.slegers@soc.kuleuven.be

**Abstract.** This paper responds to the discussion of a possible lack of reciprocity in applying ethnography in HCI research, sometimes referred to as ‘rapid ethnography’. It contributes to the discussion by describing examples of how reciprocity can be achieved when applying rapid forms of ethnography. The paper suggests five approaches for HCI researchers to aim for reciprocity while doing research: (1) making participation fun for informants, (2) giving informants a voice, (3) allowing informants to exhibit their skills or strengths, (4) offering practical help and (5) providing self-knowledge. Each of these approaches comes with some risks, which are also explained in the paper. Reciprocity should be taken into consideration from the initial start of the research projects.

**Keywords:** Ethnography · Rapid ethnography · Reciprocity

## 1 Introduction

Ethnographic methods have become a common approach in HCI research since the computer became a mainstream device in the workplace and in everyday life [1]. In HCI research, the focus shifted increasingly towards the context of users or informants (i.e. the second wave). This context gradually changed: increasingly, homes or other places besides the workplace became of interest (i.e. the third wave). Ethnography became a prime research method in HCI research, since its holistic focus on groups or cultures (including their terrain or habitat) seemed a useful approach for HCI researchers to map the context of use of the technology they were working on. Many HCI research projects use at least some techniques finding their origin in ethnography (e.g. observations, interviews). However, there are fundamental differences between ethnography as it is commonly used in HCI, and traditional ethnography. Anderson [2] for instance, refers to a misunderstanding amongst HCI professionals, who tend to see ethnography as a method for data collection, while ethnographers rather see their methods as a form of analytic reportage. He suggests that it is not ethnography that designers need, but rather field experience to better understand the context of use. Dourish [3] continues to analyze this methodological view on ethnography in HCI. He emphasizes the lack of a perspectival view, which he finds critical to what ethnography is, as researchers in HCI try to

be a passive instrument when using ethnographic methods to provide an objective representation of a setting. Of specific interest for this paper, is the idea of doing ‘rapid ethnography’, as coined by Millen [4].

Millen points out another problem with the use of ethnography in HCI, namely, the mismatch between ethnography’s demand to spend (much) time in the field and the fast pace of research in HCI. Therefore, Millen suggests the use of the term ‘rapid ethnography’ to refer to the timesaving research methods commonly advocated in HCI ethnography. In his article, Millen describes the key ideas on which rapid ethnography is based: it departs from (1) narrowing down the focus of the field research, zooming in on important activities and using key informants; (2) using multiple interactive observation techniques to increase the likelihood of discovering exceptional and useful user behavior; and (3) using collaborative and computerized iterative data analyze methods.

While these ideas are widely accepted within HCI research, the use of these ‘time deepening strategies’ is not without its critics. A recent CHI paper by Brereton et al. [5] pointed out a particular challenge for researchers doing rapid ethnography. In this paper, the authors discussed the risk of a lack of reciprocity in a rapid ethnography approach: “Rapid forms of ethnography found in design research, run a particular risk of taking without giving back to communities and rushing to quick and possibly ill-conceived design approaches” [3, p. 1183]. They express concerns about a lack of attention for culturally appropriate methods for engagement and for ensuring that participation has a clear benefit for informants.

As HCI researchers, we identify with many of the struggles Brereton et al. describe, and agree that such issues should be discussed more in the CHI community. What we would especially like to discuss further is the authors’ suggestion that HCI researchers should focus on engagement and reciprocity first, to ensure valid outcomes and avoid ill-conceived design solutions. While we agree on the importance of reciprocity in research, we also think that it is possible to focus on engagement and reciprocity while following a rapid ethnographic approach design. In our rapid ethnographic research, we – and we assume HCI researchers in general – often depart from the principles of participatory design in which mutual learning, equality between researchers and participants, and reciprocity are core values [6]. In this paper, we explore how reciprocity can be achieved in rapid ethnography and to what extent this requires HCI researchers to adjust their practices.

The contribution of this paper lies in the continuation of the discussion opened by Brereton et al. It argues how reciprocity can be a focal point in rapid ethnography, which is now so common in HCI research. We do this by describing how we, in our own experiences with rapid ethnography, explicitly try to give back to our informants. We realize that the insights and experiences that we describe in this paper are subjective and personal by nature. The story may lack systematic research for some of our assumptions and suggestions. But we see this paper as a contribution to the ongoing discussion that might inspire other researchers.

## 2 Reciprocity in Rapid Ethnography

When informants, as experts of their own experience [7], give us insight into their lives, we as HCI researchers do feel the moral need to reciprocate this ‘gift’. In line with the seminal writings of Mauss on ‘the gift’ [8], we understand the need for reciprocity as a requirement to achieve an alliance with other individuals or groups in order to build ‘solidarity’. Reciprocity is not without self-interest, but it does hold a concern for others. When considering Graebers typology of reciprocity [9], it is clear that HCI researchers should avoid a situation of ‘closed reciprocity’, where the relationship resembles a form of market exchange. Here, both parties are individualistic and try to maximize their gains. This lack of solidarity is likely to result in a lack of mutual learning and, hence, ill-fitting design solutions. Instead, we should strive towards the ideal of ‘open’ reciprocity, or a relationship of mutual commitment, where no accounts are kept. This is also considered to be an essential element in participatory design research, where a hybrid space is created between researchers and informants with little emphasis on authority and individualism [6]. This can result in more engaged informants and researchers, and ultimately, in better research. As such, there is both a moral and an epistemological component to reciprocity in HCI research: researchers are morally obligated to give back to their informants, while it may also improve their work.

Mutual commitment is often not only the goal of us, researchers. We notice that many of our informants regard their participation in our research similarly. For instance, although providing informants with a financial incentive is quite common in many research domains, we noticed that most our informants do not see financial gain as a main driver for their participation. Our informants often do not expect financial incentives; some even considered incentives as inappropriate. E.g. some informants participate in research, simply for the reason of making a contribution to science. By receiving a monetary incentive, their act of benevolence is turned into an act of monetary gain. Authors such as Sandel [10] have written extensively on how monetary gifts change the meaning of actions.

Therefore, while we certainly take incentives into account when recruiting informants, we try to look for other ways to ensure reciprocity in our research. Doing so does not necessarily require activities in addition to the original research (as Brereton et al. [5] seem to suggest). Rather, we believe that it is possible to organize the research activities themselves in such a manner that they are not only meaningful for researchers but also for the informants. Below, we list a number of such ways and give examples of how reciprocity may be attained in rapid ethnography approaches.

### 2.1 Making Participation Fun

Fun and play are important parts of human activity [11]. In our opinion, aiming for pleasure during research can make a big difference in the experience of the informants and be a form of reciprocity in itself. When informants see their participation as a fun, enriching experience, they might see the activity itself as something that has been given to them.

To illustrate this, we refer to the approach of a workshop organized recently for a project on information provision for train travellers [12]. These workshops were

organized next to a series of observations and interviews. For the workshops, a board game was developed. On the game board, a train journey was visualized. The informants were divided into teams, and were asked to compete against each other to be the first to reach the final destination of the journey. During this journey they were confronted with several incidents and asked to express their need for information when facing such an incident in real life.

While the board game was very simple (the main rules were similar to those of The Game of the Goose and The Game of Life), this approach resulted a playful and relaxed atmosphere. An evaluation of the approach showed that informants felt like they were caught by (pleasant) surprise when the method was introduced to them. The evaluation survey illustrated that the majority of the informants enjoyed the workshop. They found it to be pleasantly different from previous research activities they were involved in. While all informants had received a monetary incentive as well, due to the fun nature of the workshops, it seemed that they regarded their involvement in the research as a valuable experience in itself.

*Risks.* Making informants' involvement more fun is usually more time consuming than traditional research setup. Also, there will most likely be at least one informant who does not enjoy the supposedly pleasant activity (which was also the case in the workshops with the board games). And worse, too much focus on 'fun' during methodological development might cause the researchers to loose focus and diverting their attention from their actual research questions.

## 2.2 Giving Informants a Voice

Ultimately, our informants' contributions should lead to the development of a new product or service. But as Brereton et al. [5] also point out, this cannot always be guaranteed. In fact, it is probably more likely that research valorization is less immediate and will rather be found in academic dissemination or in long-term insights for industrial partners involved. However, this does not necessarily frustrate informants. Often, it seems that informants are grateful to be given a voice, and find the opportunity to be heard valuable in itself. Often, informants have mentioned feelings of relief, thanking the interviewer for the attention and having enjoyed sharing their story.

As such, the responsibility of the researcher when interviewing cannot be overestimated. The interviewer has to ensure a gratifying conversation, and a pleasurable course of the interview, while also gathering the data required for the research. Depending on the type of interview, researchers have a topic list or a (semi-)structured questionnaire, but they should also have a flexible mindset and time schedule, be willing to deviate substantially from the original topic list and allow informants to tell the stories they would like to tell. While this is of course true for any interview, our experience is that researchers often find it difficult to substantially deviate from their topic list during interviews. While this is a standard approach in traditional ethnography, researchers doing rapid ethnography might at times have to allow interviews to take 30 min to a full hour longer than originally planned.

To illustrate this, we refer to research activities with informal caregivers in a project aimed at developing a platform for services for micromanagement in home care [13].

Researchers repeatedly visited informal caregivers at their homes to gain insights in their caring activities and their requirements for a platform for care coordination. Most of the caregivers were under considerable stress; some even experienced helplessness about the situation they were in. Participating in the project further heightened the pressure on their lives. However, several caregivers felt grateful for the attention they were given. The researchers did their best to make the conversations pleasurable experiences, and took the issues the informants raised at heart, often thinking along to find solutions for problems they brought up, even though they were often not directly related to the project. The interviews could theoretically be dealt with within less than an hour, but the researchers did not rush. Some interviews were very personal and emotional, touching upon deep frustrations or structural problems. Afterwards, several informants thanked the researchers involved extensively, having enjoyed the conversation.

*Risks.* The informants' expectations might differ considerably from the researchers' intentions. In one project [14], we interviewed persons with dementia and their family caregivers. As the informants were recruited via a hospital, some caregivers assumed that the interviewers would be able to give them medical advice. They would bring up medication and therapy schemes and wanted to discuss alternatives with the interviewers. Researchers should be as clear as possible about their goals and manage informants' expectations.

### 2.3 Allowing Informants to Exhibit Their Strengths or Skills

In line with the previous section is the idea of providing informants with a 'stage'. In this respect, informants are asked to do something they are really good at and enjoy showing to others. As such, the researchers becomes sort of an audience for the informants, affirming that they are good at something and endorsing their skills (cf. the teacher – student model in contextual inquiry [15]). This dynamic can raise the confidence of the informants, making them more comfortable and feel good about themselves.

In fact, in ethnographic HCI research, researchers often explicitly see their informants as 'experts of their experience' [7], reassuring them that they have something to contribute to the research (often informants are doubtful whether they have anything to contribute). Providing ways and/or tools to be at the center of attention when doing something informants are good at, or tell about something they are knowledgeable of, does not only provide us with valuable insights, it also gives something back to the them: they tend to feel better about themselves.

In the project with dementia patients mentioned above, for instance, the researchers discussed what activities the patients really liked doing, what they were really good at. This activity was then used as the basis for researcher-informant interactivity. Two researchers went over to a patient and her son to cook with them for instance, as mother and son liked doing this together, because the mother could still do many cooking tasks on her own.

*Risks.* Providing a stage for informants to show researchers what they are really good at or enjoy in some cases may result in informants slowly drifting away from the focus of the research. In such situations, it can be challenging for researchers to subtly steer their informants back to the topic of the study at hand.

## 2.4 Offering Practical Help

Perhaps the most obvious form of reciprocity is offering practical help to the informant. This does not necessarily imply that researchers have to organize activities in addition to the research activities. Rather, researchers can strive to help informants while doing research.

For instance, in the home care micromanagement project mentioned above, researchers observed a person taking care of meal deliveries for care receivers with high dependency. For a full day, this person and a researcher drove from home to deliver meals. Here, the researchers started assisting this person, getting out the car themselves to hand over the meals. Similarly, in another project [16], we also were of some assistance to teachers by taking over some of their teaching duties when organizing workshops with students during their teaching hours.

This type of practical assistance is not always feasible, especially not in highly specialized work environments (e.g. during observations in surgical theaters it was impossible to be of practical help [17]). But very often, there is an opportunity for researchers to be of some assistance. Even when such assistance is very limited, it does allow researchers to communicate their intention to strive for reciprocity and equality in the research process.

*Risks.* As always with participatory observation, the researchers should make sure that by helping they do not alter the situation in such a manner that the data obtained no longer hold any relation with the phenomena they wish to observe. Also, especially with frail target groups such as care receivers, offering help might have legal implications in case things do not go as planned.

## 2.5 Providing Self Knowledge/Mirroring

While participating in research activities, or being presented with the outcomes of a research project one participated in, informants may be presented with considerable insights into their own lives, which can be deeply gratifying. For instance, cultural probes can stimulate reflection and encourage informants to closely examine certain facets of their lives, as to be more able to understand and verbalize their experiences more precisely [18]. In one project, we asked people who had recently retired to reflect on their experiences regarding this phase in their lives. We used an extensive package of cultural probes as a sensitization for later interviews and prototyping sessions. Some informants had put so much effort in the cultural probes that they refused to leave them behind. They felt the probes had become important objects reflecting their personal experience. They wanted to keep the probes, and share what they had learned about themselves with their family.

*Risks.* Although gaining self-knowledge can be deeply gratifying, it may also confront informants with an unpleasant reality they were not yet aware of or tried to forget or ignore. Self-knowledge can be confrontational: when aiming for this type of reciprocity, researchers should carefully consider the consequences of increased self-knowledge.

### 3 Discussion and Conclusion

In this paper we discussed several ways in which we try to give back to our informants during rapid ethnography. By doing this, we aimed to show that HCI research is not always about taking without giving back, as was identified by Brereton et al. [5] as a risk. We do agree that the rapid approach to ethnography is very different from traditional ethnography. However, we also firmly believe that rapid ethnography does not exclude engagement and reciprocity. We hope that our reflections inspire HCI researchers also striving for open reciprocity, mutual commitment or equal relationships in their research. While taking up some of these strategies might require an additional effort from the researchers involved, they do not require them to organize activities in addition to those activities planned for data gathering. They can be used in combination with the time deepening strategies as formulated by Millen, without having to resort to a traditional ethnographic study.

We fully agree with Brereton et al. that reciprocity should be considered first and foremost when working with informants, both for moral and epistemological reasons. Therefore, we plead for other researchers in the HCI community to take reciprocity into consideration from the start of any research project. We also think that reciprocity can be achieved in numerous small ways without having to set up additional activities, of which we gave a number of examples in this paper.

We realize that our ways of giving back to our informants may not always be very concrete or tangible. However, we find it important that HCI researchers share their attempts at reciprocity, as we did in this paper. By sharing such experiences, and by continuing the discussion about ethnography in HCI, we believe that it is possible to reduce the risks that are related to rapid ethnography, and to further improve the quality of ethnographic research in HCI.

## References

1. Crabtree, D.A., Rouncefield, D.M., Tolmie, D.P.: Ethnography and systems design. In: *Doing Design Ethnography*, pp. 7–19. Springer, London (2012)
2. Anderson, R.J.: Representations and Requirements: the value of ethnography in system design. *Hum. Comput. Interact.* **9**, 151–182 (1994)
3. Dourish, P.: Implications for design. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 541–550. ACM, New York (2006)
4. Millen, D.R.: Rapid ethnography: time deepening strategies for HCI field research. In: *Proceedings of the 3rd Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques*, pp. 280–286. ACM, New York (2000)
5. Brereton, M., Roe, P., Schroeter, R., Lee Hong, A.: Beyond ethnography: engagement and reciprocity as foundations for design research out here. In: *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*, pp. 1183–1186. ACM, New York (2014)
6. Muller, M.J., Druin, A.: Participatory design: the third space in HCI. In: Jacko, J.A., Sears, A. (eds.) *The Human-Computer Interaction Handbook*, pp. 1051–1068. Lawrence Erlbaum Associates Inc., Hillsdale (2003)

7. Visser, F.S., Stappers, P.J., van der Lugt, R., Sanders, E.B.-N.: Contextmapping: experiences from practice. *CoDesign* **1**, 119–149 (2005)
8. Mauss, M.: *The Gift: The Form and Reason for Exchange in Archaic Societies*. W. W. Norton & Company, New York (2000)
9. Graeber, D.: *Toward an Anthropological Theory of Value: The False Coin of Our Own Dreams*. Palgrave Macmillan, Basingstoke (2001)
10. Sandel, M.: *What Money Can't Buy*. Penguin, London (2013)
11. Huizinga, J.: *Homo Ludens: A Study of the Play-Element in Culture*. Beacon Press, Boston (1971)
12. Slegers, K., Ruelens, S., Vissers, J., Duysburgh, P.: Using game principles in UX research: a board game for eliciting future user needs. In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 1225–1228. ACM, New York (2015)
13. iMinds: O'CareCloudS. Organizing home care using a cloud-based platform (2012). <http://www.iminds.be/en/projects/2014/04/07/ocareclouds>
14. Slegers, K., Wilkinson, A., Hendriks, N.: Active collaboration in healthcare design: participatory design to develop a dementia care app. In: *Proceedings of Extended Abstracts on Human Factors in Computing Systems, CHI 2013*, pp. 475–480. ACM, New York (2013)
15. Beyer, H., Holtzblatt, K.: *Contextual Design: Defining Customer-Centered Systems*. Elsevier, Amsterdam (1997)
16. Slegers, K., Duysburgh, P., Jacobs, A.: Research methods for involving hearing impaired children in IT innovation. In: *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*, pp. 781–784. ACM, New York (2010)
17. Duysburgh, P., Elprama, S.A., Jacobs, A.: Exploring the social-technological gap in telesurgery: collaboration within distributed or teams. In: *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing*, pp. 1537–1548. ACM, New York (2014)
18. Gaver, B., Dunne, T., Pacenti, E.: Design: cultural probes. *Interactions* **6**, 21–29 (1999)

# Testing the Unknown – Value of Usability Testing for Complex Professional Systems Development

Kimmo Tarkkanen<sup>(✉)</sup>, Ville Harkke, and Pekka Reijonen

Information Systems Science, University of Turku, Turku, Finland  
`{kimmo.tarkkanen,ville.harkke}@utu.fi`

**Abstract.** To make an impact on the design in usability testing, the test tasks are essential ingredients for the early system development process. Complex design problems are not solved by focusing on the details of a prototype and setting the scope on what is already known by the design team. Instead, the design value of usability testing is increased by deliberately relinquishing the assumptions made and implemented into a design. In the development of complex systems, usability testing with extended scope and open-ended structure, as presented in this paper with three empirical cases, delivers not only specific knowledge about the user interactions with the system, but reveals issues that, despite rigorous user research efforts, have been overlooked in the preceding phases of system development. Therefore, we suggest applying open-ended usability test tasks for testing systems in complex settings such as in the development of health care systems.

**Keywords:** Usability testing · Test task · Design · Complex systems · Health care

## 1 Introduction

Professional systems in healthcare are designed to support work that can be described as complex problem solving [1]. Complex problem solving is characterized by the unpredictability of the process, as the path to solving any given problem may differ from the path of another. Professional autonomy and the nature of activities in the health care domain introduce many kinds of varying work practices and essential workarounds [2]. This natural unpredictability and complexity of the health care domain have a profound effect on designing and testing of health care systems, making the design of an optimal professional system a wicked problem (see [3]). As solutions to wicked problems are tested in practical settings [4], the obvious path to better health care systems is user involvement and user studies.

The method collection known as usability testing has retained its popularity as one way of validating the proposed design solutions within the modern system development methodologies. Even when leaning on close communication with the customer (e.g. agile methods) there are clear benefits to be realized by conducting user research [5]. In its classical form, usability testing is focused on detecting the usability problems of the software product and recommending correspondent changes to the design. The problem-centric approach of usability testing, the validity and the reliability of found problems

as well as the value of succeeding design recommendations are all questioned in the past [6–10]. Major challenges have been that most of the reported usability problems only confirm earlier impressions of developers [11] i.e. developers are not very interested in usability problems, nor do they react on these [9].

Therefore, the modern formative usability testing aims to influence the design process and the designed artifact with more cooperative manner with the development team than before [6–8]. However, put into the context of complex problem domains, the scope and focus of usability tests are too often traditional and narrow, not aiming at reviewing users' actual work in these contexts [12]. In a complex domain, software developers should search multiple and alternative contexts of use [1], explore the right design direction by generating and testing ideas, instead of trying to get the first design right [13]. Thus, to have an impact on design, usability tests will have to mirror the complexity of the problem domain in the planning and aim at revealing issues that bring developers closer to a solution to the wicked problem. In practice, this means questioning all that is known in the design – testing the unknown – where the focus is set on acquiring user knowledge for the development with scope that covers not only the design artefact, but the whole spheres of contexts of use and beyond.

In this paper, we introduce three usability tests conducted in the early development phases of new health care systems. Usability test tasks given to users were open-ended in order to broaden the scope and focus of tests on acquiring knowledge of contexts that would serve the design in the complex domain as well as possible. The cases introduce results and design issues that are beyond traditional. First, the results corrected designers' wrong assumptions about the current work practices at home care. Second, the results identified the adjunct roles of users in occupational health care that refined the whole scope of the development project. Third, the results manifested the low power of end-users and software developers in the development of the national health archive. After discussing the findings in relation to usability testing procedures and objectives in systems development, we conclude that usability tests can influence complex systems development in multiple and unexpected ways. We maintain that in supporting the early stages of a design process, usability test tasks have a major role in refining the focus and scope of results.

## 2 Exploring the Unknown in Systems Development

Complexity, in systems development, can be regarded as characterizing either the context of system use or the designers' knowledge of it. The former is a natural part of the domain and hard to eliminate, yet often needs to be supported or partially resolved by technological artifacts. According to Mirel [1] such complexity is different in kind, not just degree, from well-structured simple tasks. Complexity from the latter perspective often characterizes the early phases of the system development. Traditional requirements engineering is based on the assumption that it is possible to recognise and plan for static requirements before actual development and implementation work is started. Methods for conducting this type of engineering are well documented and validated and range from interviews and observations to prototyping in various forms [14]. In more

advanced software development methodologies known e.g. as agile methods, the importance of formal planning is diminished and more direct communications between the actors are preferred [15]. It is admitted that the beginning of the design process is fuzzy [16], and the requirements are allowed to evolve in ongoing participation with future users.

Knowledge of users is hardly sufficient in terms of quantity and quality in many development projects. It is, however, impossible, even a design fallacy, to completely collect such knowledge [17]. Since the work by [18], it is known that the work practices during the system deployment may not correspond to the descriptions of work created and built into the system during the implementation process. The systems are exposed to drifting in usage and objectives [19], because users discover affordances [20], new and heterogeneous uses [21], and apply workarounds in order to get their work done [22]. Even ethnographic inquiries, which can address such intricacy of specific user contexts and work practices, are ineffective and insufficient in fulfilling the needs of design [17].

While the user involvement in design has become a truism for many IS development projects [23], the original, idealistic, picture of equal power of stakeholders in participatory design [24] has been reduced in user-centered methods to a power of system developer [25]. System developers and evaluators decide how much user participation is allowed [26] and what it means that a system is well-designed [25]. Thus, the user participation is institutionalized under the logic of technology development [25]. Likewise, the user practices are planned by the design before the actual use [27]. A true participation requires that potential users and stakeholders have a possibility to formulate and express questions and problems of the design and eventually have power to define the target of the development and possible solutions to it. An example of how to achieve this is to apply simple mock-ups in tests. These will evoke a variety of comments by users around the context, whereas with more detailed mock-ups the conversation focus is on the artifacts [28]. Discrepancies found in such simple solutions are inherently valuable in opening up the design decision-making and leading to an open-ended design process that extends into actual artifact use (cf. [27]). In contrast, a predefined scope of the development project and fixed interests of the developers will fundamentally ignore other possible perspectives and unforeseen parts of the solution.

## 2.1 The Role of Usability Testing and Test Tasks in the Design Exploration

Usability testing is a one of the most applied, evaluator-led, user inquiry methods under the umbrella of user-centered design. The method is well-established and seamlessly integrative with modern software development methodologies, yet regarded as a more confirming and disproving than exploratory and innovative method for designing ideas [29]. One of the reasons is that the evaluator rigidly controls what is asked, seen and provided – in good and bad [30]. This is manifested especially in the test tasks that are chosen by the evaluator and define in advance what type of results the test will produce. Therefore the tasks and task formulations are critical in collecting usability data and both focusing the attention of evaluators and setting the scope for the whole test and its possible results. To succeed in the

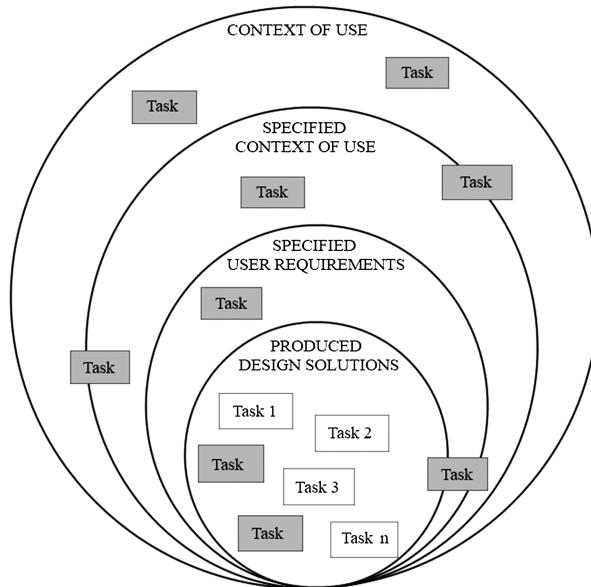
evaluation, evaluators need extensive domain knowledge [31]. Tasks and test scenarios are created based on knowledge of the domain and the product, its objectives, its target users and their supposed activities – knowledge, acquired in the previous development phases and requirements elicitation processes. A set of created tasks, as real and meaningful as possible, is further reduced based on various criteria (e.g. supposed frequency and criticality at actual use). In fact, domain knowledge is so important that domain experts without any usability expertise found more severe usability problems than usability experts in a study by [32]. Unsurprisingly, longitudinal user observations in the field revealed that testing the usability of system properties in the laboratory premises is not valid for real use situations [33]. The context-sensitivity of usability work is frequently ignored in the complex health care domain too [34], where end-users' lack abilities to contribute to system development and the wide range of IT tools in the clinical context is overlooked [35].

Since the idea of UCD in the 80's [36] various method collections, method combinations and method modifications have emerged. Usability testing has also been harnessed to study users and practices at complex work domains. For example, [12] suggests conducting studies in the field of users, exploiting multiple evaluators, building simulations, developing situation awareness assessments, implementing unattended long-term data capture and using cued retrospective think aloud method with users. Another example is to apply a 'cooperative' usability test in order to gain knowledge about the work domain [37]. The essence of the method is retrospective interpretation phases, which ask why the user acted in a certain way and thus utilize user's knowledge of the work-domain to identify and understand usability problems [37]. In the field of health care, [38] combine an interaction sequence analysis to their usability study, pre-exploring the work practices of users before the tests with a contextual inquiry [39].

The weakness of classic usability testing is that test tasks may only concentrate on (1) features of the proposed design and (2) how these can be operated (3) in the known work tasks of the users the product is supposed to support. This leads to a premature commitment to the defined requirements [40], which rules out a wide spectrum of user requirements that might be needed [41]. Conventional usability test tasks, which include clear endings and correct answers, are not applicable to the analysis of complex systems [12]. Complex systems introduce usually much higher level goals than applied in typical usability testing tasks, which may be hard to specify beforehand while lower level usability testing may result in an easy-to-use solution for the wrong set of requirements [12].

Another explanation for this problem may be the existence of different and overlapping definitions of usability [42]. At one end, usability stops where utility begins [43], on the other end, usability is not a design issue at all, but inherent and inseparable from work and other goal-oriented action of people using tools (cf. [44]). In the latter perspective, one can be interested in how the system operates, yet subordinate the system operations to work processes that eventually determine usability and utility of the system. That is not only about asking how well the system does (efficiency), but does it do the right thing at all (effectiveness). Usability studies with complex tasks have overlooked to measure the effectiveness aspect of usability [45]. Unfortunately, it may be easier to plan test tasks that are based on the current design than such that question it.

In order to increase their value the usability tests within design processes call for test tasks that fall outside the design solution and even outside the pre-specified context of use (Fig. 1). The expected benefit for the development process would be questioning and testing the design decisions made earlier (in Fig. 1 the unknown and known specified issues are presented as the nested circles of UCD process). That is bringing into discussion user needs and requirements, actual use situations and specific work practices that have not been included in the design solution and the intended sphere of usage. To define these tasks in practice, the potential user can, for example, bring her own tasks to the test session [41]. The benefit is that in addition to evaluating usability from the perspective of the design, the users' perspective is taken into consideration. Some of the user-defined tasks [41] may fall inside the proposed design solution, i.e. product functions support their execution while some may fall outside i.e. the tasks have not been implemented into the design.



**Fig. 1.** Two different sets of test tasks: Pre-planned (1-N) and user-based (shaded tasks), which may fall anywhere between the context of use and the proposed solution (nested circles).

A modification of user-defined tasks, open-ended tasks, introduces another approach to empower users in a test session. In addition, such tasks cut the link with the assumptions made in the design process. An open-ended task approach does not point to any product function but to a whole work (process) inside the sphere of the context of use. It is formulated as a high-level request for test participants to perform their work with the system under evaluation as a support. In this way, users can follow their natural flow of work and articulate their situational needs more freely [46]. Thus, instead of pre-defined models of work, users' situated work practices act as a starting point for usability evaluation and analysis.

Next, we discuss results that were acquired with open-ended tasks in three usability tests in the health care domain. The complexity of health care work is shortly explained and further discussed in the case-specific context.

### 3 Testing Usability in Three Complex Health Care Settings

Clinical work in the health care domain involves natural complexities [2, 47, 48]. Mirel [1, 49] relates these to complex problem solving, which is characterized by, among other things, vague goals, multiple methods and paths and lack of a distinct right answer. Handling these complexities of health care requires boundary crossing, polycontextuality and horizontal expertise [50]. For example, physicians face new problems in patient interventions that cannot be quickly turned into codified and repeatable procedures. According to Berg [51], clinical work is “*...characterized by the constant emergence of contingencies that require ad hoc and pragmatic responses. Although much work follows routinized paths, the complexity of health care organizations and the never fully predictable nature of patients' reactions to interventions, result in an ongoing stream of sudden events. These have to be dealt with on the spot, by whomever happens to be present, and with whatever resources happen to be at hand...*

” Furthermore, the work is characterized by distributed decision making, by ‘multiple viewpoints’ and by its ‘inconsistent and evolving knowledge bases’ [51]. The organization of health work mostly involves multiple stakeholders, the goals and preferences of whom may not be aligned [51].

#### 3.1 Highlighting the Drifting Work Practices in Home Care

Home care involves many professionals from distinct disciplines who work in a cooperative and coordinated manner to provide care services for people living in their own homes. Home care workers describe the work as “*highly personalized caring labor that often seeps out of its formal boundaries into informal, unpaid activities*” [52]. In addition, the complexity of the domain is highlighted by the emergence of sudden and unpredictable care situations and needs due to less frequent monitoring of the patients.

In this domain, a mobile application was designed for nurses to be used during home care visits to clients. In order to perform patient visits, the nurses need access to client’s contact information (where to go), care plans (what to do), and possibility to view earlier care treatments and actions (what has been done). This work process as well as an analysis of stakeholders, user profiles and use cases was represented by the developer in seven pages long document describing the context of use of the mobile application. In addition, user requirements of the new application design were based on knowledge about a desktop version of the application implemented earlier into the home care organization by the same developer. The new mobile application aimed, however, to partly replace the desktop application. While the desktop application is currently used in the office premises, before and after the daily visits to inquire and entry clients’ health data, the new tablet-based mobile application aimed to offer the same features and information during the visits at clients’ premises.

A usability test for a paper prototype of the mobile application was conducted with four home care nurses working in two different units. The users were given an open-ended task as follows: "Your name is Z.Z. and you work as a nurse in the home care unit. Today is Tuesday 17th March. You are going to your third home visit of the day. The next client is called A.A. Please, perform your work and use this new application for support when needed." Due to data contained and presented in the paper prototype, the open task needed to be more detailed than originally planned i.e. the day, the name of the nurse and the name of the client have been fixed already by the paper slides and were unchangeable during the test. For the same reason, users needed to simulate their work in the test. Users were asked to think aloud while performing their client visit and documentary tasks with the artifact, and administrators intervened when necessary in order to understand the actions performed. Each session lasted 1.5 h and was video recorded and transcribed.

Apart from some more minor issues, we found that every user had unexpected difficulties finding the purpose of the visit. The reason was lack of information about daily care tasks in the application. In nurses' terms this "daily information" is the most critical information to start the working day. The current work practice is that the daily information about each patient visited is entered as a free text to the desktop application at the end of the day. Next working day it is printed on the paper, carried along the day and fulfilled with new remarks about patients. The daily information serves various purposes. First, it is to inform the nurse what care actions are needed in the specific visit while preparing for visits. Second, it serves as a to-do list for future tasks (e.g. call the daughter on Friday/bring medicine next week) and a checklist for ongoing and past visits. Third, it is used to inform and communicate to other nurses about client related work. Fourth, and most importantly, the nurses conceive patients' current health situations through these entries of daily information.

It is notable that the daily information is a combination of three distinct text fields in the desktop application, yet managed as one set of information on the paper. All three fields were not implemented into or represented as a whole in the mobile version. The mobile application design was based on the assumption that the care plan functionality would be sufficient i.e. answer what kind of treatment the client requires. While the real information need was about the client needs for one particular visit, the care plans provided rather general and stable information. However, the original assumption of developers about the information content was not totally wrong, because the use of the care plan functionality on the desktop system had drifted during the years; the plans were not updated very often because they could not be accessed during the visits. The daily information on the paper had overridden the care plan, which had somewhat outdated data.

The open task applied in the usability test allowed us to distinguish and compare three distinct sets of work practices in the analysis of observations: (a) the work practices with the current desktop application in use, (b) the planned work practices implemented in the mobile application and (c) the actual work practices with the mobile application in the test session. The results were valuable for the further design of mobile application. Not only was more knowledge acquired about users and their activities with the mobile application, but about the foundation of these activities. Based on this knowledge, more

justified design decisions could be made by both the developer and by the user organization: For example, whether it is necessary to keep the care plans updated and accessed during the visits i.e. to change the current work practices to benefit from the mobile application.

### **3.2 Identifying the Dual Roles of Users in Occupational Health Care**

The aim of the development project was to redesign a current electronic patient record system (EPR) used by occupational health care providers. The first phase of the development was confined to functionalities on the physicians' and nurses' desktop module. The redesigned system was supposed to cover the whole care process and serve as a tool for nurses and physicians to carry out their daily work tasks. The initial requirements specification by the developer determined that with the desktop module physicians and nurses could manage appointments, health record entries, health measurements (blood pressure, weight etc.), laboratory examinations, prescriptions and customer invoices. The work of physicians is highly knowledge-and information intensive. It is hardly describable as general processes although routines based on legislation and evidence-based diagnosis practices exist. Complexities stem from patient interventions that require all the possible resources, information and tools at hand irrespective of the initial conditions and assumptions.

The first actual user research was carried out by a third party company renowned and specialized in industrial design. They interviewed seven end-users and observed their work practices and problems with the current software application in use. Findings of the user research were explicated and analyzed in 15 pages long report. This knowledge of users was supplemented and refined in cooperation with the domain experts of the developer organization. As a result, the design company created an initial wireframe of the EPR application.

A usability test for a paper prototype of the EPR system was conducted by the authors for two nurses and four physicians in the premises of the care provider. The users were given an open-ended task to "perform a patient visit and use the redesigned EPR when needed" i.e. to continue their work rather normally from the next patient as the sessions were arranged in the middle of the work day. However, the data in the paper prototype could not allow treating real patients during the test. Users were asked to think aloud while performing their work with the artifact and administrators intervened when necessary. Each session lasted about an hour and was video recorded and transcribed.

The most interesting finding in the tests was what the test subjects did after the patient visit. The nurses that participated in the test were operating the system very well, basically entering health data in a structured form to the system without major usability problems. However, nurses were far more interested in what kind of data they entered than how it was done. The reason was that they expected to fill in data that can answer practical questions from the patients and the companies they serve. For example, a company that pays for the occupational health care may want to know "how many employees in our company have high blood pressure?" which questions the nurses are responsible to answer. Thus, we found out how the nurses not only had a role of a care worker but also an information broker and analyst, who compiles statistics for different

well-being reports, communicates these to customers and frequently answers diverse health related questions.

In a similar manner during the test sessions, the physicians were highly focused on and expected to find the system functionality related to tracking the number of patient visits per day and the status of invoicing of each patient visit. The reason was that even the physicians had dual roles in the occupational health care. Firstly, they keep a common doctor's practice and care patients coming from the customer companies of the health care provider. The redesign of the EPR system was targeted at this role of the physicians. However, the physicians are also individual entrepreneurs who have their own business, which was not considered in the redesign. Physicians' business is performed in the premises of the care organization and with the tools and infrastructure, such as the redesigned EPR module, provided by the organization. Although the paper prototype introduced the customer invoicing functionality, the physicians worried whether their personal entrepreneurial requirements are implemented into the system and considered in the development. Due to the early stage of development and the scope of the project these requirements were not visible.

The usability test with the open task allowed discovering and defining different roles of the nurses and physicians of the system. While these roles were not discovered earlier or deliberately ignored by the development team, the user requirements of all relevant stakeholders were not present at the system specifications, which further indicated that the project scope of redesigning the EPR system was somewhat misaligned. Thus, a critical review of design project scope was a necessary action.

### **3.3 Exemplifying the Clinical Problems and Power Relations of National Health Archiving**

The aim of the development project was to build a National Archive of Health Information, which is centralized data storage for health records that are accessed and used via local EPR systems. The basic idea was that the health data created in one local EPR system is stored into the national archive and can be later retrieved into the same or another EPR system in another health care organization. In addition, the operational logic of the archive required that health records and related management practices were standardized across the nation. The development of the archive was initiated in 2007 and the requirements specification was led by government institutions in an open and public way in cooperation with EPR developers and related stakeholders such as pharmacies and medical associations. Such a massive development project is very complex from the design point of view. Furthermore, the wide coverage of the project, i.e. the services of every health care unit and work of every health care professional whether in public or private organizations, naturally includes characteristics of a complex domain just as the health care does in general.

A usability test for a local EPR system integrated into the archive was conducted by the authors with six health care professionals (two physicians, two head nurses, a ward secretary and a home care nurse). These professionals had, among about 100 others, used the archive-integrated EPR system in real clinical work for three months' pilot period. The usability test was carried out two months after the pilot period had ended,

in order to get explicit information about the experienced problems and their causes during the pilot. To our knowledge, the pilot and the succeeding usability test, carried out during 2012, were the first attempt to test the national archive use through the local EPR system. Despite the fairly long development process actual user testing was not technically possible earlier.

In the test session, the users were given an open task to “carry out their typical work tasks using the system”, which was the same fully functional EPR system they had used earlier. Tests were arranged in the hospital premises and sessions lasted from one to two hours during which users were thinking aloud while simulating their frequent and common tasks with the EPR system. Test administrators intervened with additional questions yet no pre-defined test tasks were given to users. All tests were video recorded and transcribed.

The usability test identified extensive usability problems in concepts, vocabulary and terminology used in the national archiving. The users were familiar with professional conventions and the local agreements regarding the contents of health records, but were unable to adapt to the nationally defined standard vocabulary during the three-month pilot. The archive integrated EPR version demanded, for example, that the headings of health record entries as well as their order of appearance were nationally unified. All users experienced problems in creating record entries during the session and had experienced these during the pilot as well. The physicians had considerable difficulties in finding the latest health record entry (even their own fresh entry), which is a rather critical task and frequently performed before patient appointments in health organizations. The reason was that record entries were re-organized based on a new concept of a service event: The latest record entry could fall under an old service event and be buried in massive records. Thus, the centuries-old tradition of chronologically ordered health record was interrupted. In general, users had major difficulties in understanding how the concept of the service event should actually be applied - when a new service event should begin or end. The problem of opening and closing a service event arose because every record entry, health document, and even an act of accessing the health record, were to be handled under some specific service event. As it was a forced act by the system, the number of created service events was surprisingly high during the pilot period i.e. users bypassed the problem by creating a new service event instead of caring a patient and managing the record under an existing event. In addition to service events, the archive introduced many other new concepts (e.g. the phase of care, reason to access the record, the headings of record entries), which required radical changes to clinical work and were experienced problematic by the users in the session. For example, physicians almost lost their ability to create and read health record entries due to changes in terminology. In fact, during the pilot period the harmful effects of these new concepts were overridden by workarounds, default values and ignorance.

The usability test finally concretized the problems of national archive development experienced by the clinical practitioners. In addition to practical usability problems in the use of the EPR in clinical work, the test indicated that the clinicians lacked power in the national archive development to define concepts and rules that highly affected their daily work. The difficulties were experienced also by the local EPR developers, because they needed to follow and interpret the national system specifications, and

moreover, lean on the user research done for that part. Thus, the causes of problems were far beyond the usability of a single EPR system connected to the national archive. The problems and recommendations of the test concerned workarounds for the user organization, system changes in the limits of national specifications for the developer organization, as well as requests for the national archive developer organization to empower all relevant stakeholders in the process and abandon the concept of service events. Although the test report was praised by the user organization, clinical practitioners as well as the EPR developers, the representatives of the archive developers, i.e. the most powerful stakeholders, refused to drop service events due to over six years of development of the concept that far. However, some misunderstandings between the EPR and archive developers in translating requirements into implementation details could be pinpointed and resolved, which led to system redesigning at the both local and national ends before new implementation and pilot iteration.

## 4 Discussion

In complex work domains, the “series of short, discrete scenarios” of classic, ‘common industry format’ kind usability testing are not appropriate [12]. Thus, the question arises how to test usability within complex professional systems development so that the results are useful in steering and informing the design process. Our approach with open-ended tasks provides a relatively low-cost solution to this problem. This was demonstrated in the complex field of clinical health care where the scope of usability work needs to be broadened [35].

Empirical tests with open-ended tasks can produce various results that benefit the design and development of health care systems. What characterizes these findings is that the issues found are essentially outside the sphere of the expected or specified use cases and contexts of use. In the first case, the findings indicated lack of knowledge of developers about users’ drifting work practices and workarounds. In the second case, findings about users’ adjacent and unrecognized work roles indicated a need to critically review and refine the whole project scope. In the third case, the new national standardization of patient information and clinical practices exemplified the low power of end-users and system developer in a nation-wide development project. In consequence, apart from suggestions about the designs themselves, even domain knowledge, design project scope and user organizations’ practices could be brought into discussion.

These results are representations of different types of misfits that are frequently confronted in the use of organizational systems. For example, our case with the national data repository is a clear case of an imposed structure that causes issues on the fundamental levels of system and organizational design [53]. The case in home care found functionality misfits [54], which lead to reduced utility and efficiency while role-based misfits, which imply mismatches between responsibility and authority [54], were present in the occupational health care development. As such misfits often have their roots in the deep structures of the system-organization interaction [54] and are hard to explore even with ethnographic methods, the open-ended testing approach appears very appealing in terms of efficiency and effectiveness. We assume that such extensive results

could not have been found by testing tasks and scenarios built on the pre-specified assumptions of the user requirements and use contexts. This does, naturally, not mean that using the other methods of requirements elicitation and user involvement could be substituted by simple usability testing, but that the usability testing method can with benefit be used to validate and refocus the results of the other methods.

Practicing open-task testing is not only about posing one open question in the beginning and listening to users for the rest of the test, but requires an active role of the administrator. Of course, the fundamentals of the think-aloud technique apply to open-task testing also – people are different and they have effects on results and procedures. Perhaps the main strength in the procedure is that the administrator in open-task testing needs to learn what actions and operations users' work activities consist of. It is not only the relationship of humans and computers in interaction that become analyzed with open tasks, but the whole activity system [55], where equally relevant elements and targets of evaluation with actors and artifacts are the actions [44]. However, as Mirel [49] points out, complex work is not supported only by emphasizing actions but studying interactions between conditions, constraints and actions. Open-task test results are firmly tied to studying such activity systems, because users do, or simulate doing, their ordinary work actions involving real objectives and motivation. Therefore, questions in the test session are not limited to the open task only, but as it is difficult for users in many work contexts to articulate explicitly how they work, administrator's effort is needed before, during and after the tests to make the work visible. During the test, this may mean constant intervening by the administrator especially when work actions are simulated.

Compared with ethnographic [56] or contextual inquiry methods [39] for evaluation purposes and for designing systems for complex settings, we maintain that the open-task technique is a relatively low cost due to minimal preparing phase and short interventions although a systematic comparison of costs and resources have not been made. In addition, comparison between the results of predefined and open tasks has not yet been experimented and could not be conducted in the above described cases due to their industrial nature and practical purposes. However, many of the found problems may demand further investigation with the above methods and therefore the open-task approach is for them not a competitor, but a complement. Specifically from an evaluation perspective it is a technique to catch the most profound problems of the artifact early in the development while learning more about users and use contexts.

## 5 Conclusions

Usability test tasks are essential ingredients for the early system design process. Furthermore, tasks are fundamental to usability tests to make an impact on the design. In the development of complex systems, usability testing with the extended scope and open-ended structure as presented in this paper, delivers not only specific knowledge about the user interactions with the system, but can even reveal issues that, despite rigorous user research efforts, have been overlooked in the preceding phases of system development. The approach can disentangle the evaluation from previous design assumptions and share the ideal of participatory design where users are empowered partners of the

design and evaluation. As demonstrated empirically, the approach will benefit the fuzzy and ongoing design process in the exploration of multiple and alternative contexts and future directions of early design for complex systems. The results of the presented case studies could be incorporated into the following design iterations in practice. Therefore, for usability practitioners, we suggest applying open-ended test tasks especially for testing systems in complex settings. Yet, user-initiated test tasks can be used with benefits even in other work domains and with different types of systems than discussed here (see [57]). Furthermore, we encourage technology developers and user organizations as well to acknowledge the wide spectrum of the possible outcomes of usability testing, some of which are not manageable by designers only but require attention and actions by managers at different levels and organizations. We want to maintain that the open-ended approach is not overriding the traditional type of testing with narrower scope and focus on the design solution. Instead, by increasing our understanding about the context, it also gives more credibility to such problem lists, severity ratings and design recommendations.

## References

1. Mirel, B.: Usefulness: focusing on inquiry patterns, task landscapes, and core activities. In: Mirel, B. (ed.) *Interaction Design for Complex Problem Solving*, pp. 31–63. Morgan Kaufmann, Burlington (2004)
2. Ferneley, E.H., Sobreperez, P.: Resist, comply or workaround? an examination of different facets of user engagement with information systems. *Eur. J. Inf. Syst.* **15**(4), 345–356 (2006)
3. Rittel, H.W., Webber, M.M.: Dilemmas in a general theory of planning. *Policy Sci.* **4**(2), 155–169 (1973)
4. Coyne, R.: Wicked problems revisited. *Des. Stud.* **26**(1), 5–17 (2005)
5. Gothelf, J.: *Lean UX: Applying Lean Principles to Improve User Experience*. O'Reilly Media Inc., Sebastopol (2013)
6. Hornbæk, K.: Dogmas in the assessment of usability evaluation methods. *Behav. Inf. Technol.* **29**(1), 97–111 (2010)
7. Wixon, D.: Evaluating usability methods: why the current literature fails the practitioner. *Interactions* **10**(4), 28–34 (2003)
8. Hertzum, M.: Problem prioritization in usability evaluation: from severity assessments toward impact on design. *Int. J. Hum.-Comput. Interact.* **21**(2), 125–146 (2006)
9. Molich, R., Dumas, J.S.: Comparative usability evaluation (CUE-4). *Behav. Inf. Technol.* **27**(3), 263–281 (2008)
10. Gray, W.D., Salzman, M.C.: Damaged merchandise? a review of experiments that compare usability evaluation methods. *Hum.-Comput. Interact.* **13**(3), 203–261 (1998)
11. Hornbæk, K., Frøkjær, E.: Comparing usability problems and redesign proposals as input to practical systems development. In: *Proceedings of SIGCHI Conference on Human Factors in Computing Systems*, pp. 391–400. ACM (2005)
12. Redish, J.: Expanding usability testing to evaluate complex systems. *J. Usability Stud.* **2**(3), 102–111 (2007)
13. Greenberg, S., Buxton, B.: Usability evaluation considered harmful (some of the time). In: *Proceedings of SIGCHI Conference on Human Factors in Computing Systems*, pp. 111–120. ACM (2008)

14. Paetsch, F., Eberlein, A. Maurer, F.: Requirements engineering and agile software development. In: 2012 IEEE 21st International Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises, pp. 308–308. IEEE Computer Society (2003)
15. Fowler, M., Highsmith, J.: The agile manifesto. *Softw. Dev.* **9**(8), 28–35 (2001)
16. Sanders, E.B.-N., Stappers, P.J.: Co-creation and the new landscapes of design. *Co-design* **4**(1), 5–18 (2008)
17. Stewart, J., Williams, R.: The wrong trousers? beyond the design fallacy: social learning and the user. In: *Handbook of Critical Information Systems Research*, pp. 195–237 (2005)
18. Suchman, L.A.: *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge University Press, Cambridge (1987)
19. Ciborra, C.: *From Control to Drift: The Dynamics of Corporate Information Infrastructures*. Oxford University Press, Oxford (2000)
20. Norman, D.A.: *The Psychology of Everyday Things*. Basic Books, New York (1988)
21. Salovaara, A., Öörni, A., Sokura, B.: Heterogeneous use for multiple purposes: a point of concern to is use models' validity. In: Pennarola, F., Becker, J., Baskerville, R., Chau, M. (eds.) *Proceedings of ICIS 2013* (2013)
22. Gasser, L.: The Integration of Computing and Routine Work. *ACM Trans. Inf. Syst. (TOIS)* **4**(3), 205–225 (1986)
23. Bødker, S.: When second wave hci meets third wave challenges. In: *Proceedings of the 4th Nordic Conference on Human-Computer Interaction*, pp. 1–8. ACM (2006)
24. Greenbaum, J., Kyng, M.: *Design at Work: Cooperative Design of Computer Systems*. CRC Press, Boca Raton (1991)
25. Spinuzzi, C.: A scandinavian challenge, a us response: methodological assumptions in scandinavian and US prototyping approaches. In: *Proceedings of SIGDOC*, pp. 208–215. ACM Press (2002)
26. Steen, M.: The Fragility of Human-Centred Design. Ph.D. Thesis Delft University of Technology (2008)
27. Redström, J.: RE: definitions of use. *Des. Stud.* **29**(4), 410–423 (2008)
28. Brandt, E.: How tangible mock-ups support design collaboration. *Know. Techn. Pol.* **20**, 179–192 (2007)
29. Hanington, B.: Methods in the making: a perspective on the state of human research in design. *Des. Issues* **19**(4), 9–18 (2003). MIT Press
30. Hertzum, M., Molich, R., Jacobsen, N.E.: What you get is what you see: revisiting the evaluator effect in usability tests. *Behav. Inf. Technol.* **33**(2), 144–162 (2014)
31. Chilana, P.K., Wobbrock, J.O., Ko, A.J.: Understanding usability practices in complex domains. In: *Proceedings of SIGCHI Conference on Human Factors in Computing Systems*, pp. 2337–2346. ACM (2010)
32. Følstad, A.: Work-domain experts as evaluators: usability inspection of domain-specific work-support systems. *Int. J. Hum.-Comput. Interact.* **22**(3), 217–245 (2007)
33. Riemer, K., Vehring, N.: It's not a property! exploring the sociomateriality of software usability. In: *Proceedings of International Conference on Information Systems (ICIS)*, pp. 1–19 (2010)
34. Yen, P.-Y., Bakken, S.: Review of health information technology usability study methodologies. *J. Am. Med. Inform. Assoc.* **19**(3), 413–422 (2012)
35. Kaipio, J.: *Usability in Healthcare: Overcoming the Mismatch Between Information Systems and Clinical Work*. Aalto University Publication series Doctoral Dissertations, 105/2011 (2011)
36. Norman, D.A., Draper, S.W.: *User Centered System Design: New Perspectives on Human-Computer Interaction*. L. Erlbaum Associates Inc., Hillsdale (1986)

37. Følstad, A., Hornbaek, K.: Work-domain knowledge in usability evaluation: experiences with cooperative usability testing. *J. Syst. Softw.* **83**(11), 2019–2030 (2010)
38. Viitanen, J., Nieminen, M.: Usability evaluation of digital dictation procedure - an interaction analysis approach. In: Holzinger, A., Simonic, K.-M. (eds.) *USAB 2011. LNCS*, vol. 7058, pp. 133–149. Springer, Heidelberg (2011)
39. Beyer, H., Holtzblatt, K.: Contextual design. *Interactions* **6**(1), 32–42 (1999)
40. Diaper, D.: Scenarios and Task Analysis. *Interact. Comput.* **14**(4), 379–395 (2002)
41. Cordes, R.E.: Task-selection bias: a case for user-defined tasks. *Int. J. Hum.-Comput. Interact.* **13**(4), 411–419 (2001)
42. Alonso-Ríos, D., Vázquez-García, A., Mosqueira-Rey, E., Moret-Bonillo, V.: Usability: a critical analysis and a taxonomy. *Int. J. Hum.-Comput. Interact.* **26**(1), 53–74 (2009)
43. Nielsen, J., Landauer, T.K.: A mathematical model of the finding of usability problems. In: *Proceedings of the INTERACT 1993 and CHI 1993 Conference on Human Factors in Computing Systems*, pp. 206–213. ACM (1993)
44. Nurminen, M.I., Reijonen, P., Vuorenheimo, J.: *Tietojärjestelmän Organisatorinen Käyttöönotto: Kokemuksia ja Suuntavivoja* (Organizational Implementation of IS: Experiences and Guidelines). Turku Municipal Health Department Series A (2002)
45. Frøkjær, E., Hertzum, M., Hornbæk, K.: Measuring usability: are effectiveness, efficiency, and satisfaction really correlated? In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 345–352. ACM (2000)
46. Tarkkanen, K., Reijonen, P., Tétard, F., Harkke, V.: Back to user-centered usability testing. In: Holzinger, A., Ziefle, M., Hitz, M., Debevc, M. (eds.) *SouthCHI 2013. LNCS*, vol. 7946, pp. 91–106. Springer, Heidelberg (2013)
47. Plsek, P.E., Greenhalgh, T.: The challenge of complexity in health care. *Br. Med. J.* **323**(7313), 625–628 (2001)
48. Rouse, W.B.: Health care as a complex adaptive system: implications for design and management. *Bridge* **38**(1), 17–25 (2008). National Academy of Engineering, Washington
49. Mirel, B.: Dynamic usability: designing usefulness into systems for complex tasks. In: *Content and Complexity: Information Design in Technical Communication*, pp. 233–261 (2003)
50. Engeström, Y., Engeström, R., Kärkkäinen, M.: Polycontextuality and boundary crossing in expert cognition: learning and problem solving in complex work activities. *Learn. Instruction* **5**(4), 319–336 (1995)
51. Berg, M.: Patient care information systems and health care work: a sociotechnical approach. *Int. J. Med. Inform.* **55**(2), 87–101 (1999)
52. Aronson, J., Neysmith, S.M.: “You’re not just in there to do the work” depersonalizing policies and the exploitation of home care workers’ labor. *Gend. Soc.* **10**(1), 59–77 (1996)
53. Sia, S.K., Soh, C.: An assessment of package-organisation misalignment: institutional and ontological structures. *Eur. J. Inf. Syst.* **16**(5), 568–583 (2007)
54. Strong, D.M., Volkoff, O.: Understanding organization-enterprise system fit: a path to theorizing the information technology artifact. *MIS Q.* **34**(4), 731–756 (2010)
55. Engeström, Y.: Expansive learning at work: toward an activity theoretical reconceptualization. *J. Educ. Work* **14**(1), 133–156 (2001)
56. Viller, S., Sommerville, I.: Ethnographically informed analysis for software engineers. *Int. J. Hum.-Comput. Stud.* **53**(1), 169–196 (2000)
57. Tarkkanen, K., Reijonen, P., Harkke, V., Koski, J.: Co-constructed tasks for web usability testing. In: *Proceedings of the IADIS International Conference on Interfaces and Human Computer Interaction (IHCI)*, pp. 79–86. IADIS Press (2013)

# An Empirical Investigation of Gaze Selection in Mid-Air Gestural 3D Manipulation

Eduardo Velloso<sup>1()</sup>, Jayson Turner<sup>1</sup>, Jason Alexander<sup>1</sup>,  
Andreas Bulling<sup>2</sup>, and Hans Gellersen<sup>1</sup>

<sup>1</sup> School of Computing and Communications, Infolab21, Lancaster University,  
Lancaster LA1 4WA, UK

{e.veloso, j.turner, j.alexander}@lancaster.ac.uk,  
hwg@comp.lancs.ac.uk

<sup>2</sup> Max Planck Institute for Informatics, Perceptual User Interfaces Group,  
Campus E1 4, 66123 Saarbrücken, Germany  
andreas.bulling@acm.org

**Abstract.** In this work, we investigate gaze selection in the context of mid-air hand gestural manipulation of 3D rigid bodies on monoscopic displays. We present the results of a user study with 12 participants in which we compared the performance of Gaze, a Raycasting technique (2D Cursor) and a Virtual Hand technique (3D Cursor) to select objects in two 3D mid-air interaction tasks. Also, we compared selection confirmation times for Gaze selection when selection is followed by manipulation to when it is not. Our results show that gaze selection is faster and more preferred than 2D and 3D mid-air-controlled cursors, and is particularly well suited for tasks in which users constantly switch between several objects during the manipulation. Further, selection confirmation times are longer when selection is followed by manipulation than when it is not.

**Keywords:** 3D user interfaces · Eye tracking · Mid-air gestures

## 1 Introduction

Interaction fidelity—the degree with which the actions used for a task in the UI correspond to the actions used for that task in the real world [1]—is an active topic of research in 3D user interfaces (3DUI). Interfaces based upon free-space spatial input (e.g. mid-air gestures, tilt and turn gestures, magnetic trackers, etc.) offer this fidelity for 3DUI due to their multiple degrees of freedom and high integration of dimensions of control (i.e. many degrees of freedom can be controlled simultaneously with a single movement) [2, 3]. In particular, recent advances in unobtrusive motion capture (e.g. Kinect, Leap Motion) created a renewed interest in mid-air gestures for 3DUI.

In immersive virtual reality environments and on stereoscopic displays, such interactions allow users to manipulate virtual objects using interaction metaphors that relate more closely to real world interactions, for example, by using an isometric mapping between the virtual and physical spaces, users can reach virtual objects directly where they see them. However, a large number of 3D activities, such as gaming, graphic design

and 3D modelling are still mostly conducted on conventional monoscopic desktop displays. This setup creates a discontinuity between the physical and the virtual environments, and therefore does not allow users to directly grasp objects in three dimensions. In this desktop context, common mid-air interaction techniques for 3D selection are *Raycasting* (in which the user's hand controls a 2D point that determines the direction of pointing) and the *Virtual Hand* (in which the user controls a 3D representation of his hand and makes selections by intersecting it with virtual objects) [2]. See Argelaguet et al. for a survey of selection techniques for 3D interaction [4].

As the eyes provide a natural indication of the focus of the user's interest, eye trackers have been used for pointing in a wide variety of contexts without necessarily requiring a representation on the screen, showing higher speeds than conventional techniques [5]. Even though gaze pointing for computing input has been investigated since the 80's [6], studies on gaze pointing for 3DUI started with work by Koons et al., who built a multimodal interface integrating speech, gaze and hand gestures [7]. Early work was also conducted by Tanriverdi and Jacob, who found it to be faster than an arm-extension technique with a 6DOF magnetic tracker in a VR environment [8]. Cournia et al. found conflicting results that suggest gaze is slower than a hand-based Raycasting technique with a wand [9]. These works only investigated selection tasks, but in practice, common 3D interaction tasks involve further manipulation steps after selection, such as translation and rotation. Given that gaze alone is impractical for all steps, several works combined gaze with additional modalities, but few explored the context of 3D user interfaces. In particular, when using gaze for selection and mid-air gestures for 3D manipulation, is there a cost in performance in switching modalities?

Even though gaze has been explored in a variety of multimodal configurations [10], few works explored the combination of gaze and mid-air gestures. Kosunen et al. reported preliminary results of a comparison between eye and mid-air hand pointing on large screen in a 2D task that indicate that pointing with the eyes is 29 % faster and 2.5 times more accurate than mid-air pointing [11]. Hales et al. describe a system in which discrete hand gestures issued commands to objects in the environment selected by gaze [12]. Pouke et al. investigated the combination of gaze and mid-air gestures, but in the form of a 6DOF sensor device attached to the hand [13]. They compared their technique with touch, and found that the touch-based interaction was faster and more accurate.

The conflicting results in the literature highlight the importance of further work investigating gaze selection for 3DUI, particularly considering that technical advances made eye tracking technology significantly more accurate, precise and robust than the devices and techniques used in previous works. In this work, we present an investigation of gaze selection for mid-air hand gestural manipulation of 3D rigid bodies in monoscopic displays. We conducted a study with three tasks. In the first task, we compared three 3D interaction techniques for selection and translation: a 2D cursor controlled by the hand based on Raycasting, a 3D cursor controlled by the hand analogous to a Virtual Hand and Gaze combined with mid-air gestures. In the second task, we also compared the same three techniques but in a selection and translation task involving multiple objects. In our pilot studies we found that when participants used the Gaze + Mid-Air Gestures technique, they reached out for objects even though they did not have to. We hypothesised that this action was due to the clutching required for manipulation.

To test this hypothesis, users performed a third task, in which we compared the selection time in the case where users were only required to select an object to the case where they also had to translate the object after selecting it.

Our results show that gaze selection is faster and more preferred than conventional mid-air selection techniques, particularly when users have to switch their focus between different objects. We also discovered a significant difference in the time to pinch after the object was gazed at between selection only tasks and selection followed by translation, indicating that the context of the selection impact the selection confirmation time.

## 2 Related Work

### 2.1 Human Prehension

Prehension is formally defined as “the application of functionally effective forces by the hand to an object for a task, given numerous constraints” [14], or more informally as the act of grasping or seizing. Different authors proposed ways of modelling this process. In Arbib’s model, the eyes (perceptual units), arms and hands (motor units) work together, but under distributed control to reach and grasp objects [14, 15]. The perceptual schema uses the visual input provided by the eyes to locate the object and recognise its size and orientation. The motor schema can be divided into two stages: reaching (comprised of a quick ballistic movement followed by an adjustment phase to match the object’s location) and grasping (including adjusting the finger and rotating the hand to match the object’s size and orientation, followed by the actual grasping action). Paillard’s model begins with the foveal grasping, in which the head and the eyes position themselves towards to object. Then, according to shape and positional cues, the arms and hands locate and identify the object, in open and closed loops, until finally grasping it, performing mechanical actions and sensory exploration [14, 16].

In the context of mid-air gestures for 3D user interfaces, reaching is analogous to selection and grasping to the confirmation of the selection. In this work, we investigate how human prehension can be supported in a desktop monoscopic 3D environment. In all conditions we studied, grasping (confirmation) was performed by a pinch gesture, similar to how we would grasp physical objects, but the selection step varied across conditions. The 3D cursor includes a reaching step similar to normal prehension, only offset due to the discontinuity between the virtual and physical worlds. The 2D cursor also contains a reaching step, but only in two dimensions. The Gaze condition only requires foveal grasping, as when the user looks at the object, she only needs to pinch to confirm the selection. However, as we show in the results of task 3, when the user grasps the object for further manipulation, she still reaches out for it.

### 2.2 Mid-Air Interaction for 3D Manipulation

Due to our familiarity in manipulating physical objects with our hands, a considerable effort of the HCI community has been put into developing input devices and interaction techniques that leverage our natural manual dexterity to interact with digital content.

An important interaction paradigm in 3D interaction is isomorphism: a strict, geometrical, one-to-one correspondence between hand motions in the physical and

virtual worlds [2]. Even though isomorphic techniques are shown to be more natural, they suffer from the constraints of the input device (e.g. the tracking range of the device) and of human abilities (e.g. the reach of the arm). When targets are outside the user's arm reach, techniques such as Go-Go [17] and HOMER [18] can be used to extend the length of the virtual arm [4]. Other works have explored different modalities for 3D interaction, including feet movements [19], tangible interfaces [20], and computer peripherals (e.g. 3D Connexion SpaceNavigator).

### 2.3 Gaze in Multimodal Interactions

Gaze-based interaction is known to suffer from a few challenges [21]: inaccuracy (due to the jittery nature of eye movements and technological limitations), double-role of visual observation and control, and the Midas Touch problem (the unintentional activation of functionality due to eye tracking being always-on [22]). To address these problems gaze is usually combined with other input modalities and devices.

Stellmach et al. investigated combinations of gaze with a wide variety of modalities [10], including a keyboard [23], tilt gestures [23, 24], a mouse wheel [24], touch gestures [24–26] and foot pedals [27]. A common interaction paradigm in gaze-based interaction is that of *gaze-supported interaction*—gaze suggests and the other modality confirms [25]. An example of a gaze-supported interaction technique is MAGIC pointing, which warps the mouse cursor to the area around the gaze pointing [28]. Fine positioning and selection confirmation are performed normally with the mouse.

These works have shown that multimodal gaze-based techniques are intuitive and versatile enough to work in a wide variety of contexts, ranging from small mobile devices to large public displays [29].

In this work, we have a similar goal to Stellmach and Dachselt, that of *seamless* selection and positioning [26]. Whereas in their work, they achieved this with different touch-based techniques for mobile devices, the context of 3D user interfaces requires extra degrees of freedom that are better suited for mid-air gestures.

### 2.4 Gaze and Mid-Air Gestures

Kosunen et al. reported preliminary results of a comparison between eye and mid-air hand pointing on large screen in a 2D task that indicates that pointing with the eyes is 29 % faster and 2.5 times more accurate than mid-air pointing [11]. The techniques investigated in their paper are analogous to our 2D Cursor and our Gaze technique, as they also used a pinch gesture for selection confirmation. In this paper, we extend their work to 3D manipulation, also comparing them to a 3D cursor. Also, their task involved 2D translation of objects, whereas ours involves 3D translation.

Pouke et al. investigated the combination of gaze and mid-air gestures, but in the form of a 6DOF sensor device attached to the hand [13]. Their system supported tilt, grab/switch, shake and throw gestures. They compared their technique with a touch-based one, and found that the touch-based interaction was faster and more accurate, mainly due to accuracy issues with their custom-built eye tracker. We aimed to minimise tracker accuracy problems, by using a commercial eye tracker with a gaze estimation

error of 0.4 degrees of visual angle. Our study also differs from theirs in that the mid-air gestures investigated by them were based on a tangible device, rather than hand-only gestures.

Yoo et al.'s system tracked the user's head orientation (as an approximation for the gaze point) and the 3D position of the hands for interacting with large displays [30]. Bowman et al. investigated pointing in the direction of gaze, but also approximating it to the head orientation [2]. Such approximations only work when the user is looking straight ahead. Hence they are only suitable for large scale interactions, such as with large displays and fully immersive virtual environments. In a desktop setting, the head orientation is not a good approximation for the point of regard, as the user is constantly facing the same direction.

Cha and Maier proposed a combination of gaze and mid-air gestures for a multi-display use case [31]. These authors presented architectural implementation details of their system, but did not present any evaluation results or interaction design decisions.

## 2.5 Gaze and 3D User Interfaces

Stellmach and Dachselt proposed two ways in which 3D user interfaces can benefit from eye tracking: first, understanding how users visually perceive 3D scenes can assist the design of new 3DUIs and second, eye trackers can be used for direct control of 3DUIs [32].

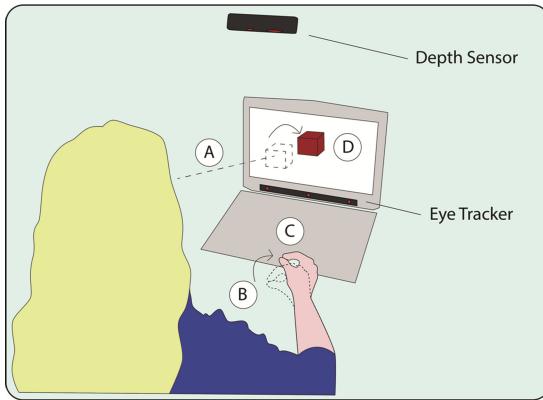
Examples of the first group of applications include studying players' gaze patterns in 3D video games to improve level design and graphics [33], to improve gaze behaviour and body animations of virtual agents [34] and to enhance the rendering of depth-of-field blur effects [35].

In the second group are applications controlled directly by gaze. In the past few years, companies such as Tobii and SMI started marketing eye trackers for the wider consumer market aimed primarily at gaming, which stimulated developers to create the first commercial gaze-enabled games [36]. The research community has also demonstrated several examples of interaction techniques and game prototypes in this context (see Sundstedt for an overview [37]).

The popularity of head-mounted displays such as the Oculus Rift created renewed interest in exploring eye tracking within Virtual Reality. Tanriverdi and Jacob compared selection time and the users' ability to recall spatial information in a VR between two techniques: the gaze position and pointing with a magnetic tracker [8]. They found that the gaze technique was significantly faster, but led to more difficulties in recalling the locations of items they had interacted with. On the other hand, Cournia et al. compared gaze and hand pointing in VR and found hand pointing to perform better [9]. Following Cournia et al.'s recommendation, we implemented our 2D cursor using raycasting, as it seemed to outperform arm extension techniques (such as the one used by Tanriverdi and Jacob [8]). These works, however, investigate 3D interaction in an immersive VR environment, whereas we use a monoscopic display. Duchowski et al. also investigated eye tracking in virtual reality in applications ranging from monitoring users' gaze for aircraft inspection training [38] to providing a visual deictic reference in collaborative environments [39].

### 3 Experimental Setup

We recruited 12 right-handed participants (6 M/6F), aged between 20 and 43 years (median = 28). Three wore glasses and one wore contact lenses in the study. Figure 1 shows our experimental setup. Participants sat in front of an 18'' laptop running a custom application built in the *Unity* game engine. Gaze was tracked at 30fps using a Tobii EyeX tracker mounted under the display, with an average gaze estimation accuracy of 0.4 degrees of visual angle. Hands were tracked using an Asus Xtion PRO LIVE sensor, with resolution of  $640 \times 480$  (30 Hz), mounted facing down on a  $0.82 \text{ m} \times 1.0 \text{ m}$  rig. Pose estimation and gesture recognition were performed using 3Gear Systems' *Nimble SDK*.



**Fig. 1.** Gaze selection for 3DUI: The user selects the object by looking at it (A), pinches (B), and moves her hand in free-space (C) to manipulate it (D).

We implemented three interaction techniques for selecting and translating objects in our 3D scene:

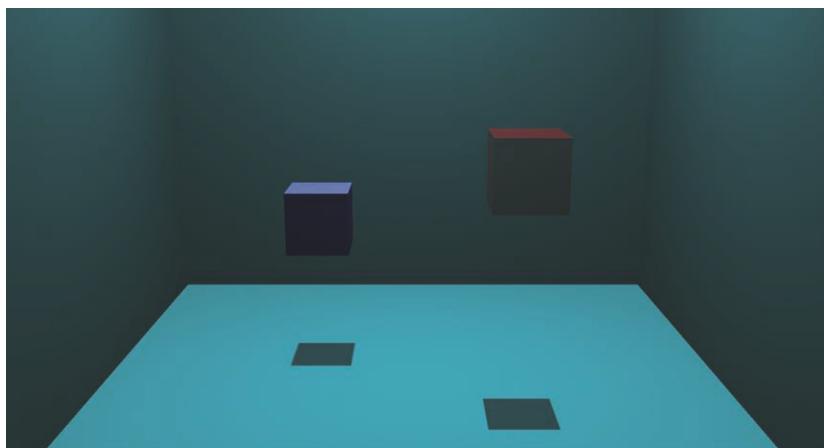
- **Gaze (Gaze-Supported Mid-Air Gestures):** the user looks at the object he wishes to select, pinch, move his hand to translate the object, and releases the pinch to disengage from the interaction.
- **2D Cursor (Raycasting):** the user moves his hand on the plane parallel to the screen (up/down and left/right), which moved a cursor on the camera plane of the scene (moving the hand towards and away from the screen had no effect on the cursor). Targets were selected by hovering over them, (similar to a mouse cursor) and pinching. Then, the user moved his hand to translate the object and released the pinch to disengage from the interaction. Note that in this interaction technique, whereas the selection step uses only the XY coordinates of the hand, the translation step uses all three (XYZ).
- **3D Cursor (Virtual Hand):** the user moves his hand around the space above the desk, which moved a sphere cursor in the virtual environment in three dimensions. Because we used an isomorphic mapping between the physical space and the 3D

scene, any movement of the hand was directly translated in an equivalent movement of the cursor. To select an object, the user intersects the sphere cursor with the desired object and pinches. The user then moves his hand to translate the object and releases the pinch to disengage from the interaction.

Upon arrival, participants completed a consent form and a demographics questionnaire. We calibrated the eye and hand trackers with the manufacturers' default procedures. Participants then performed three 3D interaction tasks, described in the following sections. After all tasks were completed, we conducted an open-ended interview about their experience in using the interaction techniques.

#### 4 Task 1: Translating a Single-Object

In Task 1, we compared completion times for two hand-based and one gaze-based selection techniques in a translation task. Participants were presented with a 3D environment containing one blue and one red cube (see Fig. 2). The task was to pick up the blue cube with a pinch gesture using each technique to select it, match its position to that of the red cube by moving their right hand whilst pinching, and drop it at the position of the red cube by releasing the pinch. When the blue cube intersected with the red cube, the red cube would turn green, indicating that the object could be released.



**Fig. 2.** Task 1: Users picked up the blue cube using each of the three techniques to select it and pinching to confirm the selection. They then moved this cube until it touched the red cube, which, in turn, would change its colour to green. The trial was concluded by releasing the pinch (Color figure online).

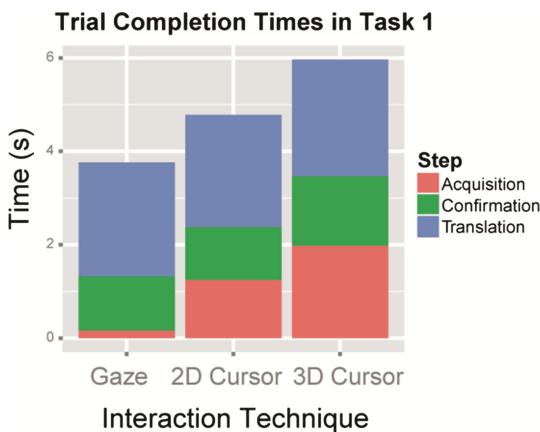
Participants performed the tasks in three blocks, each with 18 trials for each technique, in a counter-balanced order, for a total of  $3 \text{ blocks} \times 3 \text{ techniques} \times 18 \text{ trials} = 162$  interactions. In each trial, the starting position of the cubes changed, but the distance between them remained constant. In the final block, after completing all trials for each technique, participants completed a questionnaire in which they rated each technique

on a 7-point scale with respect to speed, accuracy, ease of learning and use, eye, hand and arm fatigue, intuitiveness, mental and physical effort, comfort, suitability for the task and personal preference. After completing all blocks they ranked the techniques in terms of speed, accuracy, comfort and personal preference. We discarded the first block from further analyses as a practice round.

#### 4.1 Results

We compared the mean completion times between each technique across all trials, as well as the times of each step of the task, namely the time to acquire the blue cube (Acquisition), the time to pinch to confirm the selection (Confirmation) and the time to move it to the red cube (Translation). We tested the effects of the technique on the dependent variables using a one-way repeated-measures ANOVA (Greenhouse-Geisser corrected in case Mauchly's test revealed a violation of sphericity) and post hoc pairwise t-tests (Bonferroni corrected).

The mean trial completion time using Gaze (3.76 s) was 21.3 % shorter than using the 2D Cursor (4.78 s) and 37.0 % shorter than using the 3D Cursor (5.97 s) (see Fig. 3). The effect of technique on mean completion time was significant ( $F_{2,22} = 24.5, p < .01$ ) with significant differences between all combinations of techniques ( $p < .05$ ).



**Fig. 3.** Mean task 1 completion time split by step

The Acquisition Time using Gaze (161 ms) was 87.2 % shorter than using the 2D Cursor (1.25 s), and 91.9 % shorter than using the 3D Cursor (1.98 s), with a significant effect of the technique ( $F_{2,22} = 194.5, p < .01$ ). Post hoc tests showed significant differences between all combinations of techniques at  $p < .05$ . We did not find a significant effect of the technique neither on the confirmation time ( $F_{1,2,13,2} = 3.1, p = .07$ ) nor on the translation time ( $F_{2,22} = .12, p = 0.88$ ).

In the questionnaires, Gaze received higher scores than the other two techniques along all dimensions, except for eye fatigue, for which it scored the lowest of all three

(but the difference was not statistically significant). Eleven participants ranked gaze as their preferred technique overall, with only one user preferring the 2D cursor. Nine users indicated the 3D cursor as the worst technique and three indicated the 2D cursor. A similar pattern was found for Accuracy, Speed and Comfort rankings.

## 4.2 Discussion

The results from Task 1 are in line with Tanriverdi and Jacob [8]. Even though their setup was VR-based, it seems that Gaze also outperform other 3D selection techniques in monoscopic displays. Unlike Cournia et al., Gaze also outperformed Raycasting for selection, but as suggested by these authors, Raycasting performed better than Virtual Hand [9].

Both Tanriverdi and Jacob and Cournia et al. investigated 3D selection, but not in the context of further manipulation. We also included a translation task to analyse whether the selection technique influenced the completion time of subsequent manipulation tasks (for example, by requiring clutching or adjustment of the hand position after selection). Because we found no significant difference in the confirmation and translation tasks, we cannot affirm that these interaction techniques have any effects on the manipulation task time, even though we observed certain hand clutching in the Gaze and 2D Cursor conditions. As shown in Fig. 3, the only significant cause for the difference in the task completion time was in the object acquisition.

## 5 Task 2: Sorting Multiple Objects

In Task 1, we showed that the acquisition time using gaze is significantly shorter than using the other techniques. However, Gaze is known to suffer from inaccuracies, due to the jittery nature of eye movement, calibration issues and gaze estimation error. The goals of the second task were twofold: to investigate whether eye tracking inaccuracies would impair object selection in cluttered environments and to investigate how the faster selection times enabled by gaze can speed up tasks in which the user is required to rapidly manipulate different objects in sequence. We hypothesised that, because users do not have to necessarily move their hands to pick up new objects with Gaze, the fact that they could start the manipulation from wherever their hands were would speed up switching between objects.

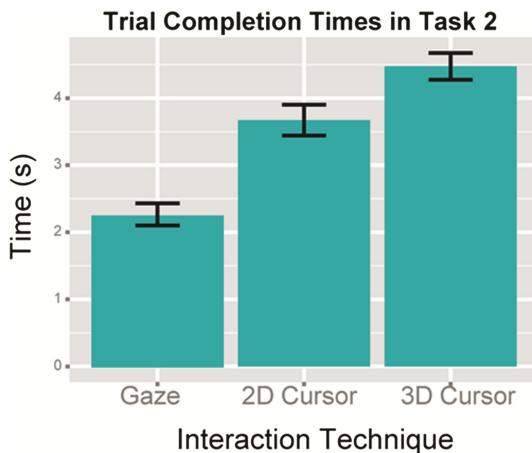
Participants were presented with the same environment, now containing six black and six white chess pieces (see Fig. 4). The right and left walls were coloured in white and black, respectively. Participants were asked to pick up each chess piece and move it to the appropriate wall. When the object collided with the corresponding wall, it disappeared. If the object collided with the wrong wall, it did not disappear, but we incremented an error counter. Each participant performed three trials with 12 pieces, totalling  $3 \text{ trials} \times 3 \text{ techniques} \times 12 \text{ pieces} = 108 \text{ interactions}$ . In the last trial, after each technique, they answered the same questionnaire as before. After all trials were completed, they completed the preference ranking questionnaire again. We discarded the first trial of each technique as a practice round.



**Fig. 4.** Task 2: Users picked up each chess piece and moved it to the appropriate side of the virtual environment.

## 5.1 Results

The mean time to put away each piece with Gaze (2.27 s) was 38.4 % shorter than with the 2D cursor (3.67 s) and 49.4 % shorter than the 3D cursor (4.47 s) (see Fig. 5). We found a significant effect of the technique on Completion Time ( $F_{2,22} = 37.7, p < .01$ ) and significant differences between all combinations at  $p < .05$ .



**Fig. 5.** Mean trial completion times in task 2. Gaze was significantly faster than the other two techniques.

The mean rate of incorrectly placed pieces with the 3D Cursor (1.92 %) was 71.3 % smaller than with the 2D cursor (6.70 %) and 82.7 % smaller than the Gaze (11.1 %). We found a significant effect on Error Rate ( $F_{2,22} = 8.19, p < .01$ ). The post hoc tests

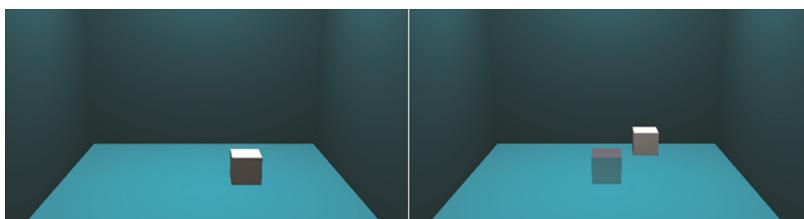
showed significant differences only between Gaze and the 3D Cursor ( $p < .05$ ). No considerable differences were found in the questionnaire responses between the first and second task.

## 5.2 Discussion

The task completion times in Task 2 were significantly shorter than in Task 1. The reason for this is that whereas the selection step required precision, the translation step did not—as soon as the object hit the correct wall, the task was complete. For the hand-based tasks this represented a similar gain in speed (30 % for the 2D Cursor and 34 % for the 3D Cursor), but a much higher gain in speed for the gaze technique (66 %). This shows that, even though it comes at a price of accuracy, Gaze is particularly well suited for tasks in which there is constant switching between different objects being manipulated. Examples of such tasks include organising furniture in architectural applications, playing speed-based 3D puzzle games and switching between different tools in a 3D modelling application.

## 6 Task 3: Selection Only vs. Selection and Manipulation

In our pilot studies, we noticed an interesting phenomenon when observing participants using gaze-assisted mid-air gestures. In the Gaze condition, once participants looked at the object, they could pinch from wherever their hands were and start manipulating the object from there. However, users still slightly reached out to the general position of the object, either to open up space for subsequent manipulations or due to a natural tendency to reach out as when handling real objects.



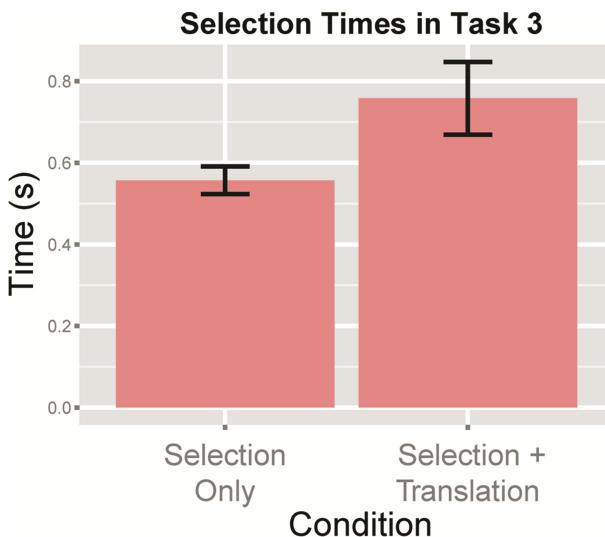
**Fig. 6.** Task 3: We compared the selection time between only selecting the object to the selection time with subsequent manipulation.

We hypothesized that this “clutching” before the translation would delay the selection confirmation when compared to selecting the object without any subsequent manipulation. To test this, we conducted a third task with the same 3D environment (see Fig. 6-A). In one condition, a white cube appeared at random positions, but always at the same Y coordinate and at one of two Z coordinates (one in the foreground and one in the background). To reset the gaze point, between each trial the cube would show up at the centre of the environment. Participants were asked to look at the cube and make a pinch gesture, after which the cube would disappear. To avoid participants predicting

the timing of the pinch gestures, we added a random delay uniformly distributed between 500 ms and 1.0 s before each trial. The second condition was similar to the first, but after pinching, the user was asked to drag the white cube to a red cube at centre of the environment (see Fig. 6-B). Participants performed three blocks, each containing 20 trials of each task (not counting the gaze-resetting steps), for a total of 3 blocks  $\times$  2 tasks  $\times$  20 trials = 120 interactions.

## 6.1 Results

We compared the time to perform the pinch gesture after having acquired the object with their gaze. The time in the Selection Only condition (557 ms) was 26.5 % shorter than in the Selection + Translation (758 ms) (see Fig. 7). A Welch's t-test revealed that this difference was significant ( $t_{11} = -2.69, p < .05$ ).



**Fig. 7.** Selection times in task 3. Users took longer to select the object when they were going to manipulate it afterwards.

## 6.2 Discussion

Our results show that the time taken to select an object with Gaze is significantly longer when the user plans on manipulating it afterwards. We offer three possible explanations for this phenomenon. First, when the user must translate the object after picking it, there is an additional planning step in the prehension process, adding some extra time for cognitive processing. Second, it is our natural behaviour to reach out in the general direction of where objects are. Third, with Gaze, even though the object can be selected from wherever the hand is, this initial position must allow enough room for the subsequent manipulation. Therefore, if the user's hand is not in an appropriate position, she

must clutch it before picking the object up. From our observations, we believe the third explanation to be the most likely one.

## 7 Discussion

The results of Task 1 show that the acquisition time varied significantly between techniques, with Gaze being the fastest, followed by the 2D Cursor. We did not find a significant modality switch latency, as once the object was acquired, participants took approximately the same time to pick it up with a pinch gesture and move it to the target. As the results from Task 2 show, the advantage of Gaze is even stronger in tasks in which multiple objects are manipulated in sequence. This gain in speed comes at the cost of accuracy, particularly in densely populated environments. Participants' opinions on the techniques also confirmed that Gaze was the most popular technique.

In task 3, we discovered a significant difference in the time to pinch after the object was gazed at between selection only tasks and selection followed by translation. Although this difference is negligible for practical purposes, it reveals an interesting aspect of human behaviour when interacting using gaze. Even though the system was calibrated so that no clutching was necessary, participants still reached out in the general direction of where the object was positioned before pinching, similarly to how they would do with physical objects. Gaze selection elegantly supports this natural behaviour. This result suggests that gaze selection should be analysed in the context of the subsequent tasks, and not as an independent phenomenon.

We conducted our experiment in a desktop environment. The presented techniques could, however be extended to standing interaction with large displays and immersive environments. Moreover, in stereoscopic displays, as the hands do not need to intersect the objects, Gaze-selection can be used without breaking the 3D illusion. Another limitation was that we only looked at translation tasks, but the same could be investigated for rotation and scaling.

Gaze-assisted mid-air manipulation allows users to select objects far away and manipulate them comfortably as if they were within reach. This allows users to rest their wrists on the desk, minimising the *Gorilla Arm* problem. This technique is also particularly useful for monoscopic displays, where the inherent discontinuity between the virtual and the physical spaces do not allow for direct manipulation and often require an extra step for positioning the cursor on the target. In fact, participants reported not having to think about this step at all and that all they had to do was to think about the object and pinch, allowing for an arguably more immersive experience and an interaction with more fidelity.

## 8 Conclusion

In this work we evaluated gaze as a modality for object selection in combination with mid-air hand gestures for manipulation in 3DUI. Whereas previous work has found conflicting results on the performance of gaze for 3D interaction, we found that gaze outperforms other mid-air selection techniques and supports users' natural behaviours

when reaching out for objects. Our findings suggest that gaze is a promising modality for 3D interaction and that it deserves further exploration in a wider variety of contexts. In particular, in future work we would like to explore how gaze can modulate the mapping between the physical and virtual environments, making it easier to reach distant objects, for example. Another avenue for investigation is how gaze can be incorporated into existing 3D applications.

## References

1. Bowman, D.A., McMahan, R.P., Ragan, E.D.: Questioning naturalism in 3D user interfaces. *Commun. ACM* **55**, 78–88 (2012)
2. Bowman, D.A., Kruijff, E., LaViola Jr, J.J., Poupyrev, I.: *3D User Interfaces: Theory and Practice*. Addison-Wesley, Boston (2004)
3. Hinckley, K., Pausch, R., Goble, J.C., Kassell, N.F.: A survey of design issues in spatial input. In: *Proceedings of the 7th Annual ACM Symposium on User Interface Software and Technology*, pp. 213–222. ACM (1994)
4. Argelaguet, F., Andujar, C.: A survey of 3D object selection techniques for virtual environments. *Comput. Graph.* **37**, 121–136 (2013)
5. Sibert, L.E., Jacob, R.J.: Evaluation of eye gaze interaction. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 281–288. ACM (2000)
6. Ware, C., Mikaelian, H.H.: An evaluation of an eye tracker as a device for computer input. In: *Proceedings of the SIGCHI/GI Conference on Human Factors in Computing Systems and Graphics Interface*, pp. 183–188. ACM, Toronto (1987)
7. Koons, D.B., Sparrell, C.J., Thorisson, K.R.: Integrating simultaneous input from speech, gaze, and hand gestures. In: Maybury, M.T. (ed.) *Intelligent Multimedia Interfaces*, pp. 257–276. American Association for Artificial Intelligence, Menlo Park (1993)
8. Tanriverdi, V., Jacob, R.J.: Interacting with eye movements in virtual environments. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 265–272. ACM (2000)
9. Cournia, N., Smith, J.D., Duchowski, A.T.: Gaze-vs. hand-based pointing in virtual environments. In: *CHI 2003 Extended Abstracts on Human Factors in Computing Systems*, pp. 772–773. ACM (2003)
10. Stellmach, S.: *Gaze-supported Multimodal Interaction* (2013). <http://www.dr.hut-verlag.de/978-3-8439-1235-8.html>
11. Kosunen, I., Jylha, A., Ahmed, I., An, C., Chech, L., Gamberini, L., Cavazza, M., Jacucci, G.: Comparing eye and gesture pointing to drag items on large screens. In: *ITS*, pp. 425–428. ACM (2013)
12. Hales, J., Rozado, D., Mardanbegi, D.: Interacting with objects in the environment by gaze and hand gestures. In: *ECEM* (2011)
13. Pouke, M., Karhu, A., Hickey, S., Arhippainen, L.: Gaze tracking and non-touch gesture based interaction method for mobile 3D virtual spaces. In: *OzCHI*, pp. 505–512. ACM (2012)
14. MacKenzie, C.L., Iberall, T.: *The Grasping Hand*. Elsevier, Amsterdam (1994)
15. Arbib, M.A.: Perceptual structures and distributed motor control. *Comprehensive Physiology* (1981)
16. Paillard, J.: Le corps situé et le corps identifié. *Rev. Méd. Suisse Romande*. 100 (1980)
17. Poupyrev, I., Billinghurst, M., Weghorst, S., Ichikawa, T.: The go-go interaction technique: non-linear mapping for direct manipulation in VR. In: *Proceedings of the 9th Annual ACM Symposium on User Interface Software and Technology*, pp. 79–80. ACM, Seattle (1996)

18. Bowman, D.A., Hodges, L.F.: An evaluation of techniques for grabbing and manipulating remote objects in immersive virtual environments. In: Proceedings of the 1997 Symposium on Interactive 3D Graphics, pp. 35–38. ACM (1997)
19. Simeone, A., Velloso, E., Alexander, J., Gellersen, H.: Feet movement in desktop 3D interaction. In: Proceedings of the 2014 IEEE Symposium on 3D User Interfaces. IEEE (2014)
20. Kitamura, Y., Itoh, Y., Kishino, F.: Real-time 3D interaction with ActiveCube. In: CHI 2001 Extended Abstracts on Human Factors in Computing Systems, pp. 355–356. ACM, Seattle (2001)
21. Stellmach, S., Dachselt, R.: Still looking: investigating seamless gaze-supported selection, positioning, and manipulation of distant targets. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 285–294. ACM (2013)
22. Jacob, R.J.: What you look at is what you get: eye movement-based interaction techniques. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 11–18. ACM (1990)
23. Stellmach, S., Stober, S., Nürnberg, A., Dachselt, R.: Designing gaze-supported multimodal interactions for the exploration of large image collections. In: Proceedings of the 1st Conference on Novel Gaze-Controlled Applications, p. 1. ACM (2011)
24. Stellmach, S., Dachselt, R.: Investigating gaze-supported multimodal pan and zoom. In: Proceedings of the Symposium on Eye Tracking Research and Applications, pp. 357–360. ACM (2012)
25. Stellmach, S., Dachselt, R.: Look & touch: gaze-supported target acquisition. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 2981–2990. ACM (2012)
26. Stellmach, S., Dachselt, R.: Still looking: Investigating seamless gaze-supported selection, positioning, and manipulation of distant targets. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 285–294. ACM (2013)
27. Göbel, F., Klamka, K., Siegel, A., Vogt, S., Stellmach, S., Dachselt, R.: Gaze-supported foot interaction in zoomable information spaces. In: CHI 2013 Extended Abstracts on Human Factors in Computing Systems, pp. 3059–3062. ACM (2013)
28. Zhai, S., Morimoto, C., Ihde, S.: Manual and gaze input cascaded (MAGIC) pointing. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 246–253. ACM (1999)
29. Turner, J., Alexander, J., Bulling, A., Schmidt, D., Gellersen, H.: Eye pull, eye push: moving objects between large screens and personal devices with gaze and touch. In: Kotzé, P., Marsden, G., Lindgaard, G., Wesson, J., Winckler, M. (eds.) INTERACT 2013, Part II. LNCS, vol. 8118, pp. 170–186. Springer, Heidelberg (2013)
30. Yoo, B., Han, J.-J., Choi, C., Yi, K., Suh, S., Park, D., Kim, C.: 3D user interface combining gaze and hand gestures for large-scale display. In: CHI 2010 Extended Abstracts on Human Factors in Computing Systems, pp. 3709–3714. ACM (2010)
31. Cha, T., Maier, S.: Eye gaze assisted human-computer interaction in a hand gesture controlled multi-display environment. In: Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction, p. 13. ACM (2012)
32. Stellmach, S., Dachselt, R.: Looking at 3D user interfaces. In: CHI 2012 Workshop on The 3rd Dimension of CHI (3DCHI): Touching and Designing 3D User Interfaces, pp. 95–98 (2012)
33. El-Nasr, M.S., Yan, S.: Visual attention in 3D video games. In: Proceedings of the 2006 ACM SIGCHI International Conference on Advances in Computer Entertainment Technology, p. 22. ACM (2006)

34. Vinayagamoorthy, V., Garau, M., Steed, A., Slater, M.: An eye gaze model for dyadic interaction in an immersive virtual environment: Practice and experience. *Comput. Graph. Forum* **23**, 1–11 (2004). Wiley Online Library
35. Hillaire, S., Lécuyer, A., Cozot, R., Casiez, G.: Using an eye-tracking system to improve camera motions and depth-of-field blur effects in virtual environments. In: Virtual Reality Conference, VR 2008. IEEE, pp. 47–50. IEEE (2008)
36. Turner, J., Velloso, E., Gellersen, H., Sundstedt, V.: EyePlay: applications for gaze in games. In: Proceedings of the First ACM SIGCHI Annual Symposium on Computer-Human Interaction in Play, pp. 465–468. ACM (2014)
37. Sundstedt, V.: Gazing at games: using eye tracking to control virtual characters. In: ACM SIGGRAPH 2010 Courses, p. 5. ACM (2010)
38. Duchowski, A.T., Shivashankaraiah, V., Rawls, T., Gramopadhye, A.K., Melloy, B.J., Kanki, B.: Binocular eye tracking in virtual reality for inspection training. In: Proceedings of the 2000 Symposium on Eye Tracking Research & Applications, pp. 89–96. ACM (2000)
39. Duchowski, A.T., Cournia, N., Cumming, B., McCallum, D., Gramopadhye, A., Greenstein, J., Sadashivan, S., Tyrrell, R.A.: Visual deictic reference in a collaborative virtual environment. In: Proceedings of the 2004 Symposium on Eye Tracking Research and Applications, pp. 35–40. ACM (2004)

# Four Eyes See More Than Two: Shared Gaze in the Car

Sandra Trösterer<sup>(✉)</sup>, Magdalena Gärtner, Martin Wuchse, Bernhard Maurer,  
Axel Baumgartner, Alexander Meschtscherjakov, and Manfred Tscheligi

Christian Doppler Laboratory “Contextual Interfaces”, Center for Human-Computer Interaction,  
University of Salzburg, Salzburg, Austria

{sandra.troesterer,magdalena.gaertner,martin.wuchse,  
bernhard.maurer,axel.baumgartner,alexander.meschtscherjakov,  
manfred.tscheligi}@sbg.ac.at

**Abstract.** Purposeful collaboration of driver and front-seat passenger can help in demanding driving situations and therefore increase safety. The characteristics of the car, as a context, limit the collaboration possibilities of the driver and front-seat passenger, though. In this paper, we present an approach that supports successful collaboration of the driver and front-seat passenger with regard to the contextual specifics. By capturing the front-seat passenger’s gaze and visualizing it for the driver, we create a collaborative space for information sharing in the car. We present the results from a study investigating the potentials of the co-driver’s gaze as means to support the driver during a navigational task. Our results confirm that the co-driver’s gaze can serve as helpful means to support the collaboration of driver and front-seat passenger in terms of perceived distraction and workload of the driver.

**Keywords:** Driving · Navigation · Collaboration · Shared gaze · Eye-tracking

## 1 Introduction

Driving a car is a demanding activity, especially in situations where the driver has to pay an increasing amount of attention to the surroundings due to, e.g., heavy traffic, bad weather or street conditions, or unfamiliarity with the region. In such demanding situations, a front-seat passenger can become a helpful source of support for the driver by, e.g., additionally monitoring the scene, providing hints, or actively guiding the driver. For example, when navigating and driving in an unfamiliar region, a front-seat passenger, who may be familiar with the region, can easily provide the driver with advice. Hence, a purposeful collaboration of the driver and the front-seat passenger can help the driver to focus on the driving task and maintain an appropriate and secure driving style (e.g., [7, 9]).

Nevertheless, the characteristics of the car as a context put limitations on this collaboration and on how information can be effectively shared between the front-seat passenger and driver. Verbal communication is easily possible because the driver and front-seat passenger are located next to each other in the very same space. The fact that

they are sitting side-by-side with head and body not turned to each other and the necessity of the driver to attend to the driving task, respectively keep the eyes on the road, are both factors that hinder natural face-to-face communication in the car (apart from verbal communication). In everyday life, however, face-to-face communication allows us to share and communicate information more easily [21]. For example, making eye-contact, following each other's gaze, monitoring what the partner is oriented toward, or showing points of interest by gesturing supports communication. In the car, these strategies are rather difficult and may even be dangerous in terms of driver distraction. Furthermore, due to the movement of the car, providing and sharing information between the driver and front-seat passenger is often more time-critical.

In our research, we aim at overcoming these potential disadvantages by providing a new way to share information between the front-seat passenger and the driver. Our approach is to capture the gaze of the front-seat passenger and visualize it for the driver. By showing the driver exactly where the front-seat passenger looks, we aim at providing a further means of communication between driver and front-seat passenger to enable a purposeful collaboration. We believe that this *shared gaze* approach could be helpful for the driver in any situation where the front-seat passenger becomes a supporter of the driver in his or her driving task. In the following, we will refer to the front-seat passenger who is actively supporting the driver in the driving task as “co-driver”. While we have already conducted an exploratory study [15] in order to validate the technical setup of our approach and identify possible future application scenarios, this paper focuses on the results of an experimental study. The main goal of the study was to explore the potentials and pitfalls of the approach, including aspects of its usefulness, caused workload, and perceived distraction of the driver.

In the following chapter, we will provide related work on gaze and collaboration in the automotive context. We then describe the shared gaze approach in detail and provide a chapter on the research goals that we targeted in our study. After describing the method and results of the study, we will discuss benefits and pitfalls of our approach and give an outlook on future work.

## 2 Related Work

### 2.1 Collaboration in the Car

The car is a social space in which the driver often interacts with other passengers. Social aspects can be considered as important influential factors when it comes to experiences occurring in the car (e.g., [10, 12]). For example, Gridling et al. [9] found that specifically during navigational tasks, drivers and front-seat passengers are often collaborating, primarily, when the output of the navigation system is misleading or confusing. They conclude that there is a potential for in-car interfaces that support this collaboration. Bryden et al. [4] investigated how passengers assist the driver's navigation task and found that collaboration was dependent on the perceived way finding abilities of the driver by both passenger and driver. They conclude that if passengers think they will be of assistance, they are more likely to help. Forlizzi et al. [7] claim that social aspects

should be considered in future navigation interfaces and recommend that these interfaces should support more interactivity in the timing and manner of information delivery. In addition, Gärtner et al. [8] report that drivers and passengers experience problems to effectively communicate with each other in the car. Verbal navigation instructions were specifically mentioned to sometimes not be efficient enough or ambiguous, e.g., when indicating a driving direction. Based this study, the initial idea of using gaze as additional means to support driver and front-seat passenger collaboration was born and realized as a first design sketch.

## 2.2 Gaze in the Car

When it comes to gaze in the car, most studies focus on the gaze of the driver, i.e., how s/he looks while driving, how s/he perceives information, or how visually distracted s/he is when performing a secondary task while driving (e.g., [5, 13]). Furthermore, eye movement research in the automotive area, to a great extent, focuses on capturing the driver's eye movements in order to detect the driver's state in terms of, e.g., inattention (e.g., [6]), in-alertness or fatigue (e.g., 2), or vigilance (e.g., [1]), while more recent research addresses the gaze of the driver as means to interact with in-vehicle information systems (IVIS, [11]).

The gaze of the front-seat passenger and its potential as further source of information for the driver is mostly neglected. Moniri et al. [17] and Moniri and Müller [18], however, present a first approach towards that direction. They claim that current IVIS do not offer any possibility for the car passengers to interact with the visible environment around the vehicle or provide any information about the visible objects in sight. Therefore, in their approach, they pursue the idea that the front seat passenger could use voice commands, such as "What is this building?", while looking at an object of interest. They compared different pointing modalities (eye gaze, head pose, pointing gesture, camera view, and view field) and found that eye gaze was the most precise modality to pick an object of interest in the visible environment of the car.

## 2.3 Gaze in Collaboration

Although gaze as means to support the collaboration of driver and co-driver has been neglected in the automotive domain, there exists plenty of research that focuses on the use of gaze when it comes to remote collaboration (i.e., natural face-to-face communication is hindered, which is quite comparable to the restraints of a car as a context). According to Brennan et al. [3], "collaboration has its benefits, but coordination has its costs" (p. 1465). They state that gaze can be a helpful means in order to reduce these costs. In their study, they found that the shared gaze was twice as fast and efficient as solitary search and even faster than shared-gaze-plus-voice in a collaborative visual search task, which is because speaking incurred substantial coordination costs.

Neider et al. [19] specifically focus on *spatial referencing* in their work, i.e., "the communication and confirmation of an object's location" (p. 718) when the collaborating

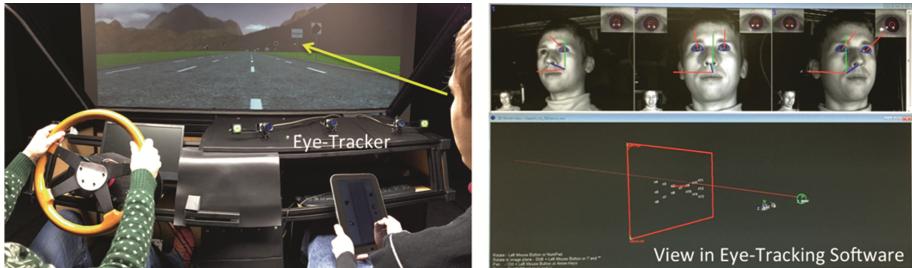
partners are remotely located. They argue that collaborative human activities often require that people attain joint attention on an object of mutual interest in a timely manner. In their study, they found that spatial referencing times (for both remotely located partners to find and agree on targets) were faster in a shared gaze condition than with shared voice, which was primarily due to faster consensus. These results suggest that sharing gaze can be more efficient than speaking when people collaborate on tasks requiring the rapid communication of spatial information.

### 3 The Shared Gaze Approach

The shared gaze approach pursues the idea of visualizing the front-seat passenger's gaze for the driver, in order to support him/her in the driving task. The aim of our approach is to provide a new way of sharing information between the driver and co-driver by using the co-driver's gaze as means to communicate spatial information more efficiently and precisely. We believe that shared gaze has potential for the automotive domain and that there are several possible application scenarios. For example, the gaze of the front-seat passenger could be used in order to help the driver in a navigational task by directly pointing to spatial reference points or to provide information about street signs that the driver may oversee in complex driving situations. Furthermore, the gaze of the co-driver could also be used to provide information regarding upcoming obstacles or dangers. On the other hand, it could also be possible to visualize the gaze of the driver for the co-driver in order to provide information for the co-driver, in situations like if the driver has already perceived certain things, e.g., hazards, landmarks, or signs. That is, we believe that our approach could be used unidirectional, but also bidirectional.

However, as already pointed out, in the automotive domain, we also have to face the challenge that working with such "visual aid" may add further distraction to the driver. Hence, we have to be careful that an expansion of the interaction space between driver and co-driver does not come with disadvantages. Therefore, we also need to carefully consider questions such as how the visualization should be exactly realized or by whom and how the visualization should be activated.

In order to study these questions we have built a prototypical implementation of the shared gaze approach in our driving simulation environment. Both the driver and co-driver are sitting in the driving simulator facing a driving simulation scenario, which is projected onto a screen in front of them. We capture the co-driver's gaze by means of an eye-tracking system and visualize it in real time as an overlay to the driving simulation projection. This enables both, the driver and the co-driver to identify where the co-driver is looking at. Additionally we have implemented a switch for the co-driver to activate and de-activate the visualization. Figure 1 shows the setup as used in our study.



**Fig. 1.** Left: Setup of shared gaze approach: driver sits behind the steering wheel; co-driver sits on the front seat; eye-tracker in front of the co-driver catches the co-driver’s gaze, which is projected along with the driving simulation onto a screen in front of the car. The yellow arrow points at the visualization of the co-driver’s gaze, which is a yellow dot. Right: Video streams from the eye-tracking cameras (top) and a model representing the co-driver’s gaze (bottom) (Color figure online).

## 4 Research Goals

Our main aim is to show the potential of our approach for the automotive domain and to gain insights in how this new approach is generally perceived in terms of its usefulness by the driver and co-driver. Furthermore, we want to find out which impact the visualization of the co-driver’s gaze has on the driver’s driving performance, the perceived distraction, and the perceived workload of the driver and co-driver. We aim at answering the question of whether gaze-supported advice by the co-driver during a navigational task provides an advantage (i.e., better driving performance, less distraction and less workload) compared to solely verbal advice of the co-driver or a solitary condition, where the driver performs the task alone.

Based on the findings from related work [3, 7, 9, 17, 19], we assume that collaborative navigation should lead to better driving performance, less perceived visual distraction, and less workload of the driver compared to solitary navigation. Furthermore, shared-gaze collaborative navigation should lead to better driving performance, less distraction, and less workload of the driver compared to solely verbal collaborative navigation. In order to investigate our research questions, we set up and conducted an explorative user study in our driving simulator. The methodological approach and the outcome of the study are described in the following section.

## 5 Study Setup

### 5.1 Participants

In total, 34 subjects (17 driver/co-driver pairs) participated in our study. One pair was excluded from data analysis due to some problems during the conduction of the experiment. The remaining 16 pairs consisted of 9 male and 7 female pairs. The participants

had a mean age of 30 years ( $SD = 6.85$ ), with the youngest subject being 19 and the eldest 48 years. The subjects were not familiar with each other and the pair assignment was based on the same gender and, if possible, same age in order to reduce stereotypical influences on the collaboration. All subjects spoke the same mother tongue and possessed a driving license. The mean mileage was 9,270 km per year. Seventeen subjects indicated that they usually drive with one passenger, six with two passengers, and three with more than two passengers. Except for one subject, all participants have been using a navigation system at least once. Almost all participants ( $n = 29$ ) have supported a driver during a navigational task in the car at least once, or have been supported by a co-driver when driving ( $n = 21$ ).

## 5.2 Experimental Design

The study was realized as permuted *within*-design with *kind of collaboration* as independent variable, consisting of four conditions: (1) **solitary**: the driver performs the navigational task alone while the co-driver sits at the front seat idly, meaning no collaboration between driver and co-driver takes place; (2) **verbal**: the driver performs the navigational task based on verbal advice provided by the co-driver; (3) **gaze**: the co-driver provides verbal advice and his/her gaze is permanently visualized for the driver; and (4) **gaze activation**: the co-driver provides verbal advice and s/he can decide when to show the driver his/her gaze (i.e., the co-driver switches the gaze visualization on or off). As dependent variables, we used *driving performance*, *perceived workload* and *perceived distraction* of the driver. Furthermore, we were interested in general impressions of the participants about the different conditions.

In order to investigate our research questions in the driving simulator, we developed a navigational task that allowed us to easily induce ambiguity of spatial reference points and to compare the different conditions in a controlled way. The development of the task followed the basics of the Lane Change Task (LCT, [14]), a common method for evaluating driver distraction. For our purposes, we used a simulated track consisting of five lanes. At both sides of the track, different street signs were shown in a random order (maximum 8 visible at once), which contained either abstract, monochrome symbols (ambiguous information, i.e., difficult to describe), or street names (unambiguous information, i.e., easy to describe). The ambiguity of signs was varied in order to mimic reality, where spatial reference points may be more or less distinct to describe as well. For example, in reality, it might be rather easy to refer to a street or exit sign compared to a certain building as reference point where the driver should make a turn.

The main task for the driver was to change into a specific lane as soon as s/he has reached a specific sign along the track while driving with constant preset speed (60 km/h), i.e., the question of “where to go” in the real world is mapped to “which lane to change to” in the task. The information of at which sign the lane change had to be performed was provided via a tablet in order to be able to provide the navigation information for both, the driver and the co-driver in the same way. On this tablet, the numbered lanes (1 to 5) and upcoming signs were visualized as plan view and an arrow was displayed abreast the specific sign, indicating the specific lane the driver has to change to. For example, as illustrated in Fig. 2, the driver would have had to change to lane 5 once

s/he has reached the sign “Steubenstraße” positioned on the left side of the track. In the verbal, gaze, and gaze activation condition, the co-driver held the tablet in his/her hands and had to guide the driver through the task based on the displayed information. In the solitary condition, the tablet was mounted in the center console and the driver had to perceive the shown information on his own. The information on the tablet was successively updated when driving along the route.



**Fig. 2.** View in the driving simulator (left) and on the tablet (right). Left: Currently the driver is driving on the middle lane of 5 lanes, i.e., lane 3. Right: The upcoming signs (either abstract monochrome symbols or street names) are shown on the tablet and the white arrow indicates that the driver needs to change to lane 5 when reaching the sign “Steubenstraße”.

### 5.3 Materials and Apparatus

**Driving Simulator.** The study was conducted in our driving simulator lab, consisting of a sitting area for participants and a car mockup situated in a projected environment (see Fig. 1). The driving simulation was realized with the software OpenDS and a data projector visualized the simulation on a  $3.28 \times 1.85$  m screen with  $1920 \times 1080$  pixel resolution.

**Task Specifications.** In total, eight different tracks were setup in OpenDS: two tracks for baseline drives, two short tracks for practicing, and four main tracks, which where assigned to the different conditions in permuted order for each pair of subjects. For the baseline drives, the lane-change information was displayed directly on the signs. In the main tracks, 16 groups of signs with eight signs per group (i.e., 128 signs in total) were shown along the five-lane road. Within a group of signs one lane change had to be performed, i.e., 16 lane changes in total. The lane change had to be performed at different positions within the sign group (e.g., at the first sign, third sign, or eighth sign), which was also permuted. Besides, the signs of one sign group were randomly arranged on the left and right side of the track. A trigger point was set randomly 10–20 m after the last sign of each group. If the driver passed this trigger, the view on the tablet was updated and showed the next sign group, which was between 140–150 m away from that position.

The distance between signs varied between 10–20 m. The short tracks were setup similarly, but only four lane changes (within four sign groups) had to be made. As regards the signs, we had a pool of eight abstract and 16 street name signs.

**Eye Tracking and Gaze Visualization.** For capturing the co-driver's gaze, we used the remote 60 Hz SmartEye Pro system. Three cameras were installed in front of the co-driver in order to properly capture head and eye movements (see Fig. 1). In order to gain as high gaze accuracy as possible, we used a 15-point calibration template that was shown on the screen during calibration. The x and y coordinates of the co-driver's gaze were captured in real time and sent to the driving simulator software, where the gaze was visualized as a yellow dot with a size of  $11 \times 11$  pixel. Beforehand, a low pass filter was applied to the eye tracking data in order to smooth the visualization of the gaze.

**Measurement Instruments.** In order to capture the driver's driving performance, his/her driving data was logged into the simulator software. In the gaze activation condition, the times when gaze was activated were automatically logged as well. Perceived workload was captured using the Driving Activity Load Index (DALI, [20]). We chose this questionnaire as it is tailored to the driving task and allows to capture different dimensions of workload. We further generated several questionnaire items to capture perceived distraction and further impressions, and a pre-questionnaire and interview guideline for a semi-structured final interview were set up in paper form.

## 5.4 Procedure

During the whole investigation, two experimenters were present. One experimenter guided the subjects through the procedure, while the other was responsible for the technical supervision.

Subjects were welcomed, introduced to each other, and seated at a desk. After providing some general information about the study, they were asked to sign an informed consent and had to fill in the pre-questionnaire. After that, the first experimenter allotted the role of the driver, respectively co-driver, and accompanied the participants to the driving simulator, where they took place according to their role. Participants could then adjust their seats, were asked to buckle up, and the appropriate acquisition of the co-drivers' head and eyes by the eye-tracking system was ensured. In order to allow the driver to get familiar with the driving simulation, s/he was asked to perform a practice drive first. If s/he felt comfortable with the driving task, s/he was asked to drive again for a defined interval (first baseline drive). After that, the eye tracking calibration of the co-driver was conducted.

Then, the first experimenter, who was sitting in the back of the car for the duration of the experiment, provided general information about the navigational task to the driver and co-driver and instructed them according to the upcoming condition. Conditions were permuted for each pair of subjects. While the driver was instructed to perform the lane change once s/he has reached the sign, the co-driver was instructed to provide the information at which sign the driver has to change the lane as forward-looking and concrete as possible. They were also allowed to exchange opinions about how to perform the task

together. For each condition, they then had the possibility to perform a practice drive in order to get familiar with the condition, which could be repeated if necessary. Then, the main navigational task followed and when it was finished, driver and co-driver were asked to fill-in questionnaires that were tailored to the condition. This procedure was repeated for each condition. The solitary condition differed slightly, i.e., the co-driver was instructed to not assist the driver and only the driver had to fill in the questionnaires afterwards. After all conditions were finished, the driver was asked to perform the second baseline drive. Finally, both participants were accompanied back to the desk where a final semi-structured interview was conducted with them both together. At the end, they were compensated with 20 Euro each and thanked for their participation. Overall, the experiment lasted about 90 min.

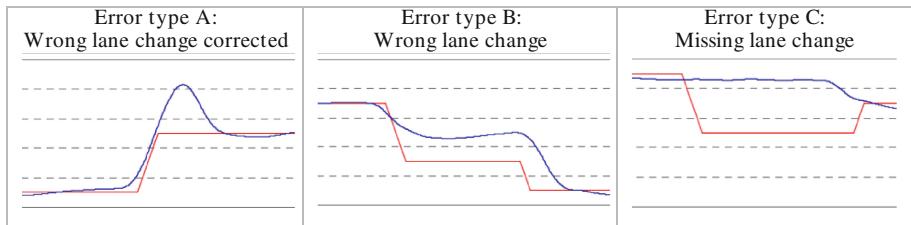
## 6 Results

All questionnaire data were preprocessed and analyzed using IBM SPSS Statistics 20. For analyzing the driving performance data we developed a tool to visualize the data and an algorithm allowing us to determine total and mean lane deviation, as well as standard deviation and variance of lane deviation. The data gained from the interviews was transcribed and Microsoft Excel was used in order to further categorize and analyze the comments of the participants according to the basics of qualitative content analysis introduced by Mayring [16].

### 6.1 Driving Performance

We first took a look at the performance of the drivers in the different conditions. In order to increase the comparability of the data, we related the driving performance of each driver with his/her performance during the baseline drives. That is, we calculated the differences between the lane deviation in the respective condition and the lane deviation in the baseline (i.e., mean of first and second drive). An ANOVA for repeated measures was calculated in order to compare the conditions. Opposed to our initial assumption that the collaborative conditions should lead to less lane deviation than the solitary condition, we could not find any significant differences between the conditions for the mean lane deviation ( $F_{(3;45)} = 2.000$ , n.s.), as well as the total lane deviation ( $F_{(3;45)} = 1.984$ , n.s.), i.e., the driving performance was comparable in all conditions.

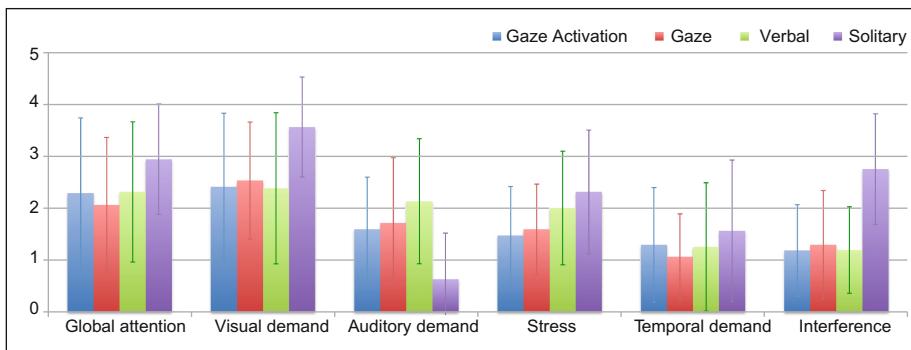
We further visually examined the driving data and found that different errors had occurred primarily during the collaborative conditions, leading to high lane deviations at certain points in time (see Fig. 3). Basically, we could identify three kinds of errors: *Error type A* - the driver changed to a wrong lane but the error was corrected before approaching the next lane change; *Error type B* - the driver changed to a wrong lane without correcting it; and *Error type C* - the driver missed the lane change. Thirteen subject pairs made at least one error during the conduction of the study. We found that most errors happened during the gaze activation condition (12), nine errors were made in the verbal condition, five in the gaze condition, and only two in the solitary condition. Hence, errors occurred more often in the collaborative conditions, but still, although almost error-free, the lane deviation in the solitary condition was comparable.



**Fig. 3.** Examples for lane change errors (the optimal lane is red, the driven lane is blue) (Color figure online)

## 6.2 Perceived Workload

In order to capture the perceived workload, we asked the drivers to fill in the DALI questionnaire [20]. Thereby, drivers had to rate their level of constraint regarding the factors global attention, visual, and auditory demand, stress, temporal demand, and interference (see Fig. 4).



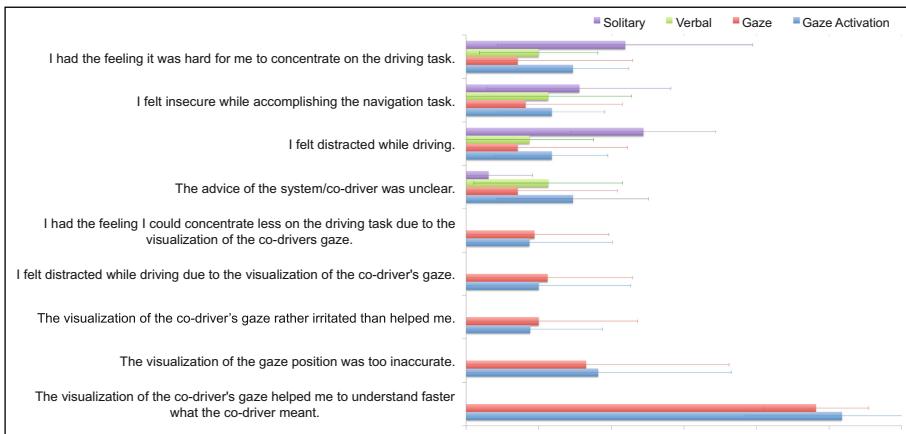
**Fig. 4.** Means and standard deviations for the different factors of the DALI questionnaire (0 = low, 5 = high)

We calculated ANOVAs for repeated measures for each factor and post-tests were calculated with Bonferroni-corrected paired t-tests. We found a main effect of condition on global attention demand ( $F_{(3;45)} = 3.731$ ,  $p < .05$ ), visual demand ( $F_{(3;45)} = 13.235$ ,  $p < .001$ ), auditory demand ( $F_{(3;45)} = 8.449$ ,  $p < .01$ ), stress ( $F_{(3;45)} = 4.189$ ,  $p < .05$ ), and interference ( $F_{(3;45)} = 13.406$ ,  $p < .001$ ). No significant differences among conditions was found for temporal demand ( $F_{(3;45)} = 1.064$ , n.s.). The post-tests revealed that in regard to global attention demand, the solitary condition led to a significantly higher global attention demand compared to the gaze condition. Furthermore, the visual demand was significantly higher compared to all collaborative conditions. The auditory demand was rated significantly lower in the solitary condition compared to the solely verbal collaborative condition. In regard to the level of stress, we could find marginally

significant differences between the solitary condition and the two gaze conditions with the latter being rated less stressful. Furthermore, the interference was rated significantly higher in the solitary condition compared to all collaborative conditions. These results suggest that the solitary condition was the most demanding with regard to nearly all DALI factors. It is further apparent that the verbal condition comes with some disadvantage due to its auditory demand. Also the induced level of stress in the verbal condition was comparable to the solitary condition.

### 6.3 Perceived Distraction and Further Interaction Qualities

In order to capture perceived distraction and further qualities of the interaction, we asked the drivers to rate several items (see Fig. 5 for an overview of drivers' ratings).



**Fig. 5.** Means and standard deviations of driver's ratings (1 = does not apply at all, 7 = does fully apply)

In regard to the issue of distraction, there was a main effect of condition ( $F_{(3;45)} = 5.736, p < .01$ ). The solitary condition was rated significantly more distractive than the verbal and gaze condition and we could not find significant differences among the collaborative conditions. We further asked whether the advice of the system or co-driver was unclear for the driver and there was also a main effect of condition ( $F_{(3;45)} = 6.033, p < .01$ ). We found that hints of the co-driver in the verbal and gaze activation condition were significantly perceived as less clear as in the solitary condition. This goes hand-in-hand with the findings from the driving data, where the ambiguity of the given advice is reflected in different lane change errors, which happened most often in the gaze activation and verbal condition.

Additionally, we asked drivers to rate a few statements specifically concerned with the gaze conditions (see Fig. 5). It is apparent that the drivers most strongly agreed that the visualization of the co-driver's gaze helped them to understand the advice of the co-driver faster. Furthermore, the mean ratings regarding a potential distraction or irritation caused

by the gaze visualization were rather low. We further asked whether the visualization of the gaze position was too inaccurate. Here, we found a slightly higher agreement, although generally still rather low.

During the study, we observed that the gaze activation condition had been somehow problematic. As it was completely left to the co-drivers when and where to show their gaze to the driver in this condition, we decided to take a closer look at the actual amount of time and frequencies the gaze was shown. Here we found a huge range among participants. Frequencies varied from 8 to 37 times (median 19) showing the gaze during the trial, while the total amount of time the gaze was shown varied from 1.9 s to 2.5 min (median 38 s). Due to this strong variation, we further examined whether the time the gaze was visible correlated with the driver's perceived distraction due to the gaze visualization. We calculated Spearman-Rho correlations and found medium correlations of  $r = 0.682$  ( $p < .01$ ) for the total time and  $r = 0.595$  ( $p < .01$ ) for the frequency of gazes and the perceived distraction of the driver. These results indicate that the more often and longer the co-driver's gaze was shown, the more distracted the driver felt.

#### 6.4 Insights from the Interviews

At the end of the study, we conducted semi-structured interviews with driver and co-driver together. We asked them about their general impressions about the different conditions, what they liked best (ranking of conditions according to their preference), and what did not work well.

There was a clear preference for the two gaze conditions among the drivers. Seven drivers preferred the gaze activation condition most and six drivers preferred the gaze condition. Only two drivers mostly preferred the verbal and one driver the solitary condition. For the co-drivers the results were comparable, although there was a stronger preference of the gaze activation condition. Ten co-drivers preferred this condition most, four co-drivers the gaze condition, and two the verbal condition.

**Drivers' Perspective.** The drivers, who voted the gaze conditions on rank one, most often stated that seeing the co-driver's gaze was helpful for them to quickly identify which sign was meant by the co-driver. Drivers preferring the gaze activation condition argued that it was good that the co-driver could use his/her gaze goal-oriented and they did not need to see the co-driver's gaze while searching for the sign. As described above, co-drivers had quite different strategies regarding the amount and way they were showing their gaze to the driver, which probably also had an impact on the drivers vote. For example, driver 11 said to the co-driver in the final interview, "I wanted to tell you that you should use your gaze more often, but it was already towards the end [end of trial]". Driver 17 mentioned it was problematic that the co-driver more often looked on the lane he had to change to rather than on the respective sign, which would have been more helpful.

Drivers preferring the permanent gaze condition stated that seeing the gaze all the time allowed them to recognize tendencies in the co-driver's gaze, e.g., if the co-driver was primarily looking at the right side of the road, they already knew that the sign must be on the right side. Five drivers also stated that they did not care whether

the gaze was shown all the time or not because they could “blank it out”, if it was not relevant for them. We also asked the participants to state which advantages and disadvantages they see in the visualization of the gaze. Almost all drivers stated that the gaze visualization made it clear for them “what was meant” and that it allowed the co-driver for giving hints more directly. As driver 16 put it, “It’s simply unmistakable [...] You are bridging communication difficulties”. Three drivers also stated that they felt relaxed and that seeing the gaze of the co-driver provided them with an additional feeling of safety.

Disadvantages of the gaze visualization were primarily mentioned with regard to permanent gaze visualization. It was stated by eight drivers that there is a potential for visual irritation, especially if the gaze is “jumping” or “going everywhere”. As stated by driver 13, “You might become irritated and with additional traffic, this is really not easy.” Three drivers meant that seeing the gaze all the time is distractive. “You only concentrate on the yellow dot” (driver 7) and two drivers felt that the visualization stressed them. Two drivers also mentioned that the gaze was a bit inaccurate.

Of the two drivers who liked the verbal condition most, driver 12 stated that “Interestingly this worked best. This was where I felt most relaxed [...] probably because we were well-rehearsed, or because the gaze point is additional information I need to mind.” The other driver (9) meant that “the descriptions became increasingly inaccurate when the gaze was added” and that, in the verbal condition, the advice has been more precise. Driver 5, who liked the solitary condition most, said that this was least stressful for her because “In the other conditions, I needed to adapt to you [the co-driver] because I would have had other terms.” However, most drivers preferred the solitary ( $n = 7$ ) and the verbal condition ( $n = 6$ ) least. Main reason was that the solitary condition was found to be distractive. In regard to the verbal advice, drivers stated that it required concentration and a precise description of the signs.

**Co-drivers’ Perspective.** Comparable to the driver’s opinion, the co-driver’s agreed that both gaze conditions were helpful because they allowed them to communicate the information faster and more distinct. However, ten co-drivers stated that seeing their gaze all the time irritated them, which is one of the main reasons they preferred the gaze activation condition. They also expressed concerns that this might be irritating for the driver as well, who probably cannot distinguish whether the gaze is meaningful right now or not. As co-driver 4 put it “It stressed me a bit when my gaze was always shown because when I did not look at a sign, I thought this could be irritating for the driver”.

Seven co-drivers explicitly expressed concerns that there have been inaccuracies in the gaze visualization. For example, co-driver 14 stated “I liked most when I could activate my gaze. This is a bit due to the fact that my gaze was not perfectly recognized and then I could try to control it a bit.” Four co-drivers stated that the gaze activation condition stressed them, though. “It was easiest when the gaze was automatically there. I did not need to decide whether I am showing him [the driver] where I look right now, but it was directly projected” (co-driver 11). In some cases, co-driver’s only noticed in the final interview that the driver would have preferred another use of gaze in the activation condition. For example, co-driver 3 commented, “For me it is totally surprising that it helped her [the driver] more when I looked at the signs. I always concentrated on

the lane, because I wanted to show her, where she has to go - how you just know it from classical navigation systems to give directions."

We also asked the co-drivers about the advantages and disadvantages of the gaze approach. Eleven co-drivers stated that the main advantage was that they could directly point at the signs and that less description was necessary. As co-driver 7 put it, "If a word just does not come to your mind, you just can look at it [the sign] and the other nevertheless knows where to go." Regarding disadvantages, seven drivers mentioned inaccuracy of the gaze, that it required them to concentrate more ( $n = 5$ ) and that it can be irritating when it is always shown ( $n = 6$ ). It was also mentioned by five co-drivers that they could not normally look around anymore in that condition, which would also become stressful over time.

**Improvements and Further Usage.** Regarding the visualization of the gaze, we found that about half of the participants (14) were completely satisfied with the size and color of the yellow dot we used in our study. The others mentioned that the color should be adaptive and providing more contrast to the surrounding (especially in the real world), or that the dot could have been a bit larger or adaptive in size as well. We further asked participants whether they could imagine using the shared gaze approach in their real car. In total 25 of 34 participants stated they would like to use or at least to try it, although they also admitted that it would be cost-dependent and rather a nice-to-have. Regarding further application areas for our approach, it was mentioned most often that the gaze of the co-driver could be used to point out hazards or dangerous situations to the driver. Others imagined that the approach could be very helpful for rescue or search operations, or in driving schools in order to teach the driver how to look correctly (e.g., while driving curves).

## 7 Discussion

The results of our study indicate that the shared gaze approach indeed comes with advantages, but also with certain limitations that need to be addressed in future research. Opposed to our initial assumption that the driving performance (in terms of lane deviation) should be better in the collaborative conditions than in the solitary condition, we found a comparable driving performance throughout all conditions. Hence, the question occurred why the collaboration of driver and co-driver did not explicitly contribute to a better driving performance. A closer examination of the data showed that it has to be taken into account that in the collaborative conditions more driving errors happened due to wrong or delayed instructions of the co-driver and that these errors of course had an impact on the overall lane deviation. From that point of view, it is actually surprising that the driving performance was still comparable to the solitary condition, where errors hardly occurred because the driver could gather the necessary information almost immediately, without processing verbal or additional visual information. Our results affirm the statement of Brennan et al. [3] that coordination has its costs, though. However, at this point, we could also see that there must be some advantage of the shared gaze approach, as the number of driving errors was least in the gaze condition in comparison to the other collaborative conditions. This is in line with our goal to support the collaboration of driver and co-driver with our approach.

The findings with regard to perceived workload and distraction were a lot more promising, though. Overall, we found that the collaborative conditions were perceived less stressful and distracting than the solitary condition in almost all cases. We found that perceived interference and visual demand were rated significantly lower in all collaborative conditions and that the stress level was experienced significantly lower in the collaborative conditions with gaze. However, the auditory demand was rated significantly higher in the verbal condition compared to the solitary condition. These results confirm our assumptions with regard to workload. In the verbal condition, the need of the driver to continuously listen to the co-drivers advice and process the verbal information obviously has disadvantages in terms of auditive demand and stress, which could be overcome with the additional visual information provided in the gaze conditions. This is in line with the findings from remote collaborative settings, where shared gaze resulted in faster and more efficient communication (e.g., [19]). We conclude that is because the driver knows earlier when and where to change the lane and, hence, stress is reduced.

Furthermore, we found that drivers rated their perceived distraction significantly lower in the verbal and gaze condition compared to the solitary condition. However, the gaze activation condition seemed a bit problematic. It turned out that the drivers experienced the hints by the co-driver least clear in the verbal and gaze activation condition. A look at the quantitative data and interview findings revealed the reasons for this. With regard to gaze activation, we found that the amount of time and frequency co-drivers showed their gaze to the driver in the gaze activation condition was varying and that different strategies were used. That is, the potential of the shared gaze approach was sometimes not fully exploited (if the gaze was used too seldom) or it was used in ways that were less helpful for the driver, e.g., by showing the lane to change and not the according sign. We also found that it sometimes also stressed the co-driver to decide whether to show his/her gaze to the driver in that condition.

We believe that this is a very important finding that has to be considered in our future research. The role of the co-driver and his/her abilities to support the driver sufficiently are apparently crucial regarding the driving performance. Subsequently, the co-driver explicitly needs to know beforehand how s/he can use his/her gaze to best support the driver, as this is obviously not as intuitive as we initially assumed. We believe that this might be also one reason why the amount of drivers favoring the gaze activation condition over the gaze condition was lower in comparison to the co-drivers. Permanent gaze visualization allowed drivers to recognize at all times where the co-driver is looking and to recognize tendencies quite early, while in the gaze activation condition drivers were more depending on the adequate use by the co-driver in order to take advantage of the gaze visualization.

Our findings further suggest that we need to take a closer look at the topic of distraction. Some driver's stated that they could "blank out" the gaze, when it was permanently visualized. However, when correlating the amount and frequency the gaze was shown in the gaze activation condition with the driver's perceived distraction, the results indicate that the more the driver sees the co-driver's gaze, the more s/he is distracted. Hence, we believe that there are probably two qualities of distraction we need to consider – distraction caused by a permanently moving object, versus distraction caused by a suddenly appearing object. Our findings suggest that both come with advantages and disadvantages. Especially from the co-driver's perspective permanent gaze visualization

is less desirable. In some cases it irritated them, but they also felt that they could not let their eyes wander around the surroundings normally anymore. In addition, it was mentioned most often by the drivers that there is a potential for irritation if the co-driver's gaze is permanently shown.

Still, almost all participants agreed that the visualization of the co-driver's gaze allowed a faster and more distinct communication of "what is meant" and the gaze conditions were preferred by most participants. Additionally, the participants could imagine further application areas of our approach, e.g., as means to point out dangerous situations to the driver.

## 8 Conclusions and Further Work

With the presented study, we aimed at identifying the potentials and pitfalls of our shared gaze approach. We found that the visualization of the co-driver's gaze has the potential to improve driver and co-driver collaboration during a navigational task and comes with less stress and perceived distraction for the driver. Although we could identify some pitfalls with regard to the gaze activation condition, we still believe that this is the condition to further work on. Based on the statements of the participants and also from a practical viewpoint, it seems feasible to show the co-driver's gaze to the driver only when it's needed. Future research, however, needs to consider the question whether giving meaningfulness to the co-driver's gaze should be an intentional decision of the co-driver or whether it could also be supported in an automated way, e.g., by just showing gazes that exceed a predefined fixation duration. Furthermore, we need to consider the question of how to visualize the co-driver's gaze in order to allow for a quick detection by the driver, especially in more complex driving environments.

In this first study, we focused on the perceived distraction of the driver, which allowed us to conclude that a basic requirement with regard to driving safety is fulfilled. However, we believe that further research is necessary regarding the actual distraction of the driver, i.e., by also capturing the driver's gaze in order to see how long and often the driver looks at the provided gaze information. This would also allow us to identify suggestions for a sound gaze visualization.

We are aware that our results are a first step for bringing the shared gaze approach into a real car. One might also ask the question, how such an approach could be technically realized in a real car. Apart from using a large head-up display, we see a lot of potential in using LED lights mounted at the bottom of the windshield as used in existing approaches that aim at providing hints about, e.g., obstacles to the driver, based on sensory information. Currently, we are preparing a study in order to investigate this kind of visualization.

In conclusion, we believe that our shared gaze approach extends the interaction space between driver and co-driver, with the potential for further application scenarios in virtue of our initial motto "four eyes see more than two".

**Acknowledgements.** The financial support by the Federal Ministry of Economy, Family and Youth, the National Foundation for Research, Technology and Development and AUDIO MOBIL

Elektronik GmbH is gratefully acknowledged (Christian Doppler Laboratory for “Contextual Interfaces”).

## References

1. Bergasa, L., Nuevo, J., Sotelo, M., Barea, R., Lopez, M.: Real-time system for monitoring driver vigilance. *IEEE Trans. Intell. Transp. Syst.* **7**(1), 63–77 (2006). IEEE Press, New York
2. Bhavya, B., Alice Josephine, R.: Intel-eye: an innovative system for accident detection, warning and prevention using image processing. *Int. J. Comput. Commun. Eng.* **2**(2), 189–193 (2013)
3. Brennan, S.E., Chen, X., Dickinson, C.A., Neider, M.B., Zelinsky, G.J.: Coordinating cognition: the costs and benefits of shared gaze during collaborative search. *Cognition* **106**(3), 1465–1477 (2008)
4. Bryden, K.J., Charlton, J., Oxley, J., Lowndes, G.: Older driver and passenger collaboration for wayfinding in unfamiliar areas. *Int. J. Behav. Dev.* **38**(4), 378–385 (2014)
5. Crundall, D.: The integration of top-down and bottom-up factors in visual search during driving. In: Underwood, G. (ed.) *Cognitive Processes in Eye Guidance*, pp. 283–302. Oxford University Press, Oxford (2005)
6. Fletcher, L., Zelinsky, A.: Driver inattention detection based on eye gaze-road event correlation. *Int. J. Robot. Res.* **28**(6), 774–801 (2009)
7. Forlizzi, J., Barley, W.C., Seder, T.: Where should I turn: moving from individual to collaborative navigation strategies to inform the interaction design of future navigation systems. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1261–1270. ACM, New York (2010)
8. Gärtnert, M., Meschtscherjakov, A., Maurer, B., Wilfinger, D., Tscheligi, M.: “Dad, stop crashing my car!”: making use of probing to inspire the design of future in-car interfaces. In: *Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI 2014)*, article 27, pp. 1–8. ACM, New York (2014)
9. Gridling, N., Meschtscherjakov, A., Tscheligi, M.: I need help!: exploring collaboration in the car. In: *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work Companion, CSCW 2012*, pp. 87–90. ACM, New York (2012)
10. Juhlin, O.: Social media on the road: mobile technologies and future traffic research. *IEEE MultiMedia* **18**(1), 8–10 (2011). IEEE Press, New York
11. Kern, D., Mahr, A., Castronovo, S., Schmidt, A., Müller, C.: Making use of drivers’ glances onto the screen for explicit gaze-based interaction. In: *Proceedings of the 2nd International Conference on Automotive User Interfaces and Interactive Vehicular Applications, AutomotiveUI 2010*, pp. 110–116. ACM, New York (2010)
12. Knobel, M., Hassenzahl, M., Lamara, M., Sattler, T., Schumann, J., Eckoldt, K., Butz A.: Clique trip: feeling related in different cars. In: *Proceedings of the Designing Interactive Systems Conference, DIS 2012*, pp. 29–37. ACM, New York (2012)
13. Land, M.F., Tatler, B.W.: *Looking and Acting: Vision and Eye Movements in Natural Behaviour*. Oxford University Press, Oxford (2009)
14. Mattes, S.: The Lane Change Task as a Tool for Driver Distraction Evaluation. IHRA-ITS Workshop on Driving Simulator Scenarios (2003). <http://www.nrd.nhtsa.dot.gov/IHRA/ITS/MATTES.pdf>

15. Maurer, B., Trösterer, S., Gärtner, M., Wuchse, M., Baumgartner, A., Meschtscherjakov, A., Wilfinger, D., Tscheligi, M.: Shared gaze in the car: towards a better driver-passenger collaboration. In: Adjunct Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI 2014), pp. 1–6. ACM, New York (2014)
16. Mayring, P.: Qualitative content analysis. In: Flick, U., von Kardorff, E., Steinke, I. (eds.) *A Companion To Qualitative Research*, pp. 266–269. SAGE Publications Ltd, London (2004)
17. Moniri, M.M., Feld, M., Müller, C.: Personalized in-vehicle information systems: building an application infrastructure for smart cars in smart spaces. In: Proceedings of the 8th International Conference on Intelligent Environments, IE 2012, pp. 379–382. IEEE, New York (2012)
18. Moniri, M.M., Müller, C.: Multimodal reference resolution for mobile spatial interaction in urban environments. In: Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, AutomotiveUI 2012, pp. 241–248. ACM, New York (2012)
19. Neider, M.B., Chen, X., Dickinson, C.A., Brennan, S.E., Zelinsky, G.J.: Coordinating spatial referencing using shared gaze. *Psychon. Bull. Rev.* **17**(5), 718–724 (2010)
20. Pauzié, A.: A method to assess the driver mental workload: the driving activity load index (DALI). *IET Intel. Transport Syst.* **2**(4), 315–322 (2008)
21. Warkentin, M.E., Luftus, S., Hightower, R.: Virtual teams versus face-to-face teams: an exploratory study of a web-based conference system. *Decis. Sci.* **28**(4), 975–996 (1997)

# Gaze+touch vs. Touch: What's the Trade-off When Using Gaze to Extend Touch to Remote Displays?

Ken Pfeuffer<sup>(✉)</sup>, Jason Alexander, and Hans Gellersen

School of Computing and Communications, InfoLab21, Lancaster University,  
Lancaster LA1 4WA, UK

{k.pfeuffer, j.alexander, h.gellersen}@lancaster.ac.uk

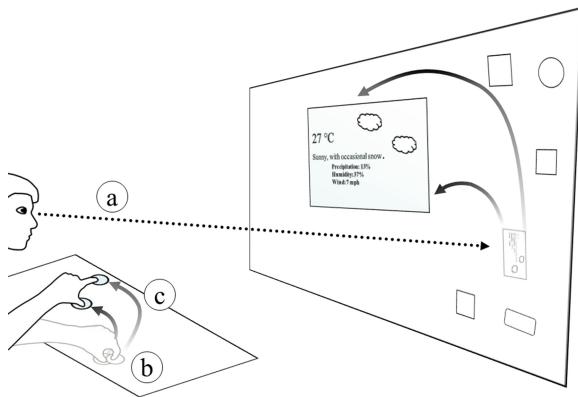
**Abstract.** Direct touch input is employed on many devices, but it is inherently restricted to displays that are reachable by the user. Gaze input as a mediator can extend touch to remote displays - using gaze for remote selection, and touch for local manipulation - but at what cost and benefit? In this paper, we investigate the potential trade-off with four experiments that empirically compare remote Gaze+touch to standard touch. Our experiments investigate dragging, rotation, and scaling tasks. Results indicate that Gaze+touch is, compared to touch, (1) equally fast and more accurate for rotation and scaling, (2) slower and less accurate for dragging, and (3) enables selection of smaller targets. Our participants confirm this trend, and are positive about the *relaxed finger placement* of Gaze+touch. Our experiments provide detailed performance characteristics to consider for the design of Gaze+touch interaction of remote displays. We further discuss insights into strengths and drawbacks in contrast to direct touch.

**Keywords:** Gaze interaction · Eye-tracking · Multitouch · Multimodal UI

## 1 Introduction

Multitouch gestures are now established on a variety of interactive surfaces such as phones, tablets, or tabletops, with much of their appeal based on the direct manipulation when touching the surface. It is of particular benefit to enable multitouch gestures on remote and large displays, as it is easy to learn, users are familiar with it, and users can easily manipulate content scattered across the large surface. However, the default touch is inherently restricted to interaction with direct surfaces in reach of the user. Although indirect interaction via a mediating mouse or touchpad can overcome this, the need for cursor dragging departs from the directness afforded by touch interaction.

A recent method to bring touch to remote displays is using the gaze modality as demonstrated in previous work ([21–25]). The control of Gaze+touch is simple: users look at the target on the remote surface for selection, and perform touch gestures on the close-proximity surface. For example, as illustrated in Fig. 1: the user looks at the target (a), touches down on the close surface (b), then performs a gesture with the same touches to manipulate the target (c). Gaze is used to determine the target at touch down, after which ‘touch’ takes over and manipulates the target.



**Fig. 1.** Gaze can be used to bring multitouch to remote surfaces: (a) look at the target, (b) touch down, and (c) perform a multi-touch gesture with the same touches.

From an input-theoretic standpoint Gaze+touch is a technique that ties characteristics of both direct and indirect touch together. Similar to direct touch, users can initiate manipulation the moment they touch down without prior cursor dragging as needed in common indirect touch techniques. Similar to indirect touch, users leverage a relaxed finger placement, use varying control-display gains, and avoid fat-finger and occlusion issues commonly associated with direct touch. For this reason, Gaze+touch should be considered as a hybrid technique rather than an instance of either direct or indirect category, which argues for a gradual characterisation of Gaze+touch through explorative evaluation as initiated in prior work [14, 21].

Yet, in light of real-world applicability, most present devices employ direct touch, making the similarities to direct touch particularly interesting. Considering a holistic perspective on ‘multitouch’, with regards to the entirety of direct manipulation rather than focusing on single gestures, Gaze+touch can be regarded as a method to bring all multitouch *as it is* to remote displays by using gaze as the mediator [21]. With this assumed, the question of trade-off becomes essential: what potential costs and benefits come with extending touch to remote displays? How does gaze selection affect the ease and familiarity of multitouch, and the quality of multi-finger and multi-hand interaction? Prior work compared Gaze+touch to head-movement based techniques [21, 22], to touch in a theoretical analysis [14], and between Gaze+touch variations for content transfer [24, 25] as well as large screens [23]. However, we are not aware of any work that empirically contrasted Gaze+touch to the default touch paradigm.

In this paper, we contribute four experiments that each compare remote Gaze+touch to standard touch interaction. The overall research question is what costs and benefits come with making touch indirect by gaze. The experiments regard task completion time, accuracy, and user feedback using the techniques. As ‘multitouch’ includes a broad range of gestures, we focus on the following commonly used gestures. Each gesture corresponds to one experiment:

1. Single-touch dragging of objects.
2. Two-touch rotation of objects across different sizes.
3. Two-touch rotation of objects of different orientations.
4. Two-touch scaling of objects.

Our experiments resulted in the following findings:

- Completion Time: Gaze+touch is as fast as touch for rotation and scaling, but slower for single-touch dragging.
- Accuracy: Gaze+touch is more accurate for rotation and scaling, but less accurate for dragging.
- Fat-Finger Problem: Gaze+touch allows to select smaller targets than touch.
- User preference: Gaze+touch is preferred for rotation and scaling, and touch is preferred for dragging.

These findings depict strength (e.g. accuracy) and shortcomings (e.g. dragging) that can be taken into consideration when designing Gaze+touch based user interfaces. While further validation is needed with real-world applicability and advanced technique design, our studies represent first groundwork in how Gaze+touch competes with and sets itself apart to standard touch interaction.

## 2 Related Work

The related work can be considered from two perspectives: research that evaluated gaze techniques, and research that evaluated touch techniques. Accordingly, we investigate the intersection of the domains by contrasting Gaze+touch to touch.

### 2.1 Evaluation of Gaze-Based Interaction

Gaze interaction is considered as fast and natural, but suffers from the Midas Touch problem (false positive activation of tasks, [9, 10, 22]). Early work therefore experimented using gaze pointing with dwell-time or button-press selection ([9, 27]). Sibert and Jacob compared gaze with dwell-time against mouse; resulting in gaze being faster than the mouse [19]. MAGIC uses gaze to replace most of the pointing of a manual device [29]. They compared MAGIC to a trackpad and found MAGIC to be faster while reducing physical effort and fatigue.

Touch received increased attention in recent time as a partner for gaze pointing ([21–25]). As touch is prime input for smartphones and tablets, it is particularly useful for interaction over distance where users point by gaze on remote displays, and confirm by touch on their local device. Stellmach and Dachselt first showed that Gaze+touch can improve selection accuracy that gaze-only usually lacks [21]. In a later work they showed that this combination also allows for dragging, scaling, and rotation of targets on distant displays [22]. In both works, Gaze+touch was compared to head based pointing techniques, and indicated performance benefit for using Gaze+touch. Turner et al. also investigated Gaze+touch with handheld and remote display, focusing on content transfer across devices ( [24, 25] ). They studied transfer techniques based on

gaze pointing with varying touch actions, showing general user acceptance and the importance of visual feedback and eye-hand coordination. Turner et al. also investigated gaze-supported rotate, scale, and translate (RST) gestures on large screens [23], indicating how subtle differences in the design of Gaze+touch techniques affect remote RST interaction. Pfeuffer et al. investigated Gaze+touch for direct, in-reach surfaces [14]. In their design space analysis, they theoretically discussed Gaze+touch in contrast to touch, arguing that using gaze can avoid typical pitfalls of touch, such as occlusion or the fat-finger issue. Our work is complementary in providing an empirical comparison of remote Gaze+touch to standard touch.

## 2.2 Comparative Studies with Direct Touch

Kin et al. compared touch on a tabletop to an indirect setting of input by mouse and output on a desktop screen [11]. The user study showed that touch can improve performance in a multi-target task, when more than one finger is used for selection. Forlines et al. also compared touch to mouse on the same tabletop device [6], indicating that touch is more appropriate for multi-finger tasks and for multiple users but less for single cases.

Cursor-based indirect touch based techniques were developed based on offset [16], multi-point [1, 3], and bimanual input [4, 28]. Potter et al. evaluated offset-based indirect touch to direct touch and found accuracy improved. Benko et al. compared indirect touch mice against direct touch [3]. They found that direct was faster as it allows explicit touch activation and resembles a single focus of interaction, contrasting the implicit cursor selection and dragging necessity. It is unclear whether these results apply to Gaze+touch since users can, like direct touch, manipulate the moment of touch down.

Other researchers proposed coupled drag gestures or consecutive touches to select small targets and increase mode selection [2, 7]. However, these techniques occupy specific touch gestures and require learning of additional techniques.

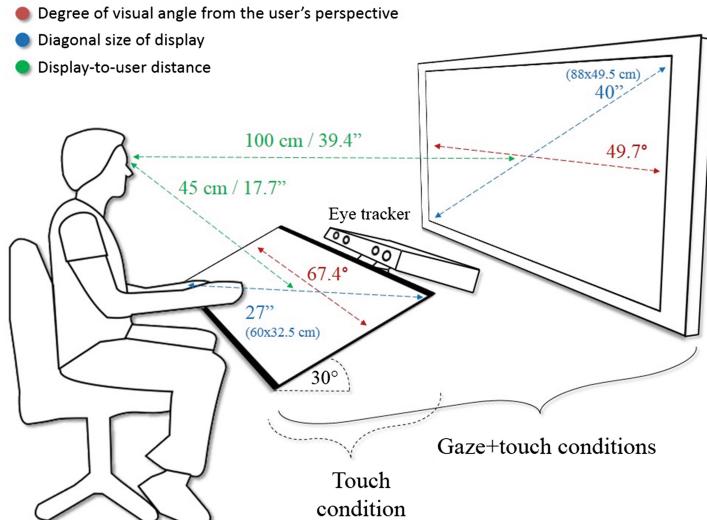
Indirect touch techniques without cursor dragging were developed by mapping touchpad and remote display 1:1 ([12, 13, 18, 26]). Schmidt et al. compared this indirect touch variant (input on horizontal surface, output on vertical surface), to direct touch (input and output on horizontal surface, [18]). Results showed lower performance with indirect touch, because users had difficulties to keep hands hovered over the surface while finding and selecting targets. With Gaze+touch, in contrast, users can touch down anywhere because selection is offloaded to gaze, eliminating the need to move and hover hands for long distances.

In summary, a variety of indirect touch techniques were developed in HCI literature and compared to direct touch as a baseline. Gaze+touch can be considered as a hybrid technique that inherits indirect characteristics without the drawbacks of cursor dragging nor absolute mappings. This makes it suited to transfer whole ‘multitouch’ to remote displays. Collectively, these reasons motivated us to take a more detailed look on this transfer, and to conduct a comparative study of Gaze+touch to touch.

### 3 Method and Design of the Experiments

**System.** The system (Fig. 2) consists of a touch display mounted at 30° close to the user (Acer T272), a large remote display (Samsung SUR40), and an eye tracker (Tobii X300). Both screens' resolution is 1080 p. The close-proximity screen supports capacitive touch input for up to ten fingers, and the eye tracker uses 60 Hz gaze tracking. The close-proximity screen is used for input and output of touch, while Gaze +touch uses it only for touch input (with visual output on the remote screen). The software runs on the SUR40 PC ( $2 \times 2.9$  GHz CPU, 4 GB RAM), implemented in Java using the MT4 J framework (<https://code.google.com/p/mt4j/>, 18/01/2015).

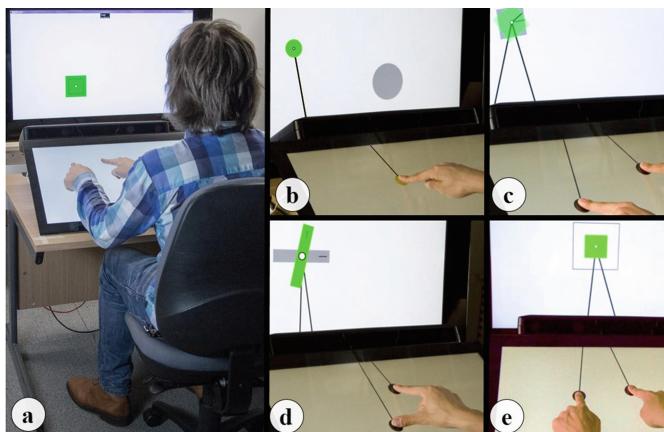
**Eye Tracking Accuracy Mechanisms and Visual Feedback.** We implemented three mechanisms to cope with issues of eye tracking hardware (e.g. imprecise tracking, eye jitter, and calibration offset, see [10, 21]). First, we calibrate each time when users look at a target, but cannot select the target at all because of inaccuracy. We use Pursuit Calibration, a calibration method more flexible than standard procedures [15]. The calibration duration was 10 s, and on average users performed two to five calibrations during the study. Second, to cope with eye jitter, we average gaze samples for 150 ms when only short eye movements occur ( $< 2^\circ$  of visual angle). Third, a gaze cursor appeared after the user fixated on a point for 1 s and did not select a target during that time. Not being able to select a target is noticeable, as usually the target is highlighted yellow when users looked it. The appearing cursor allowed the user to reposition the cursor by slight head movement if the system's gaze estimate was slightly offset preventing target selection (similar technique to Look and Lean [20]), yet allowed cursorless interaction for the majority of the time.



**Fig. 2.** Study setup (Color figure online).

**Visual Angle as a Size Metric.** To normalize measures of the close-proximity and the remote surface, we use degree of visual angle as a metric so that targets appear as the same size from the user’s view. In the Touch condition, users were approximately 45 cm in front of the screen’s center. In the Gaze+touch condition, users were approximately 100 cm from the screen (Fig. 2). Thus absolute measures have a different size from the user’s view. To normalize pixel measures, we use degree of visual angle from the user’s perspective as a metric for distance and size. For example,  $1^\circ$  of visual angle represents 25.1 px on the direct condition, and 36.2 px on the Gaze+touch condition. Figure 3 shows exemplary task designs for the Gaze+touch case:  $3^\circ$  (b),  $2^\circ \times 10^\circ$  (d), or  $4^\circ$  (c and e).

**Study Procedure.** After completing a consent form, participants were given an introduction to the study, and then conducted the four experiments. They were instructed to perform each task as quickly and as accurately as possible, with speed taking priority. Before each experiment  $\times$  technique block, the experimenter explained how the interaction technique is used, followed by a first trial with assistance. Each block was repeated five times. Notably, the whole first block was training (with assistance when necessary), and excluded from the final results. Each session lasted approximately 95 min.



**Fig. 3.** Example user setup (a) and design of each experiment: dragging (b), rotation across sizes (c), rotation across orientations (d), and scaling (e). Lines show the mapping between finger and object (not visible during study).

**Task Procedure.** Each task begins when the user touches the required number of fingers on the centre of the touchscreen, and (for Gaze+touch) looks at the centre of the remote screen. Based on a similar object dragging study [22], the object’s starting position was always placed toward a random screen corner. For rotation and scaling tasks it was  $10^\circ$  from the screen’s center; dragging tasks involved specific distances (described below). The user completed a task by pressing a button, although the actual finishing time is taken as the last time the user manipulated the object. After each

experiment  $\times$  technique block, the participant filled out a questionnaire. The categories were rated on a scale between 1 (strongly disagree) and 5 (strongly agree), and included 6 categories: ease of use, speed, accuracy, learning effort, eye tiredness, and physical effort. In addition, after each experiment users selected their preference of technique.

**Experimental design.** We used a within-subjects experimental design. Each user conducted the four experiments sequentially as presented in this paper. Within each experiment, the tested technique conditions were counterbalanced. All additional factors were randomized.

**Participants.** 16 paid participants completed the study, aged 19–31 years ( $M = 25.8$ ,  $SD = 3.3$ , 7 female). Seven wore contact lenses and none wore glasses. 15 participants owned a smartphone and were right-handed. On a scale between 1 (None) to 5 (Expert), users rated themselves as experienced with digital technology ( $M = 4.5$ ,  $SD = 0.8$ ), multitouch input ( $M = 3.4$ ,  $SD = 1.1$ ), and less experienced with eye tracking ( $M = 2.4$ ,  $SD = 1.1$ ).

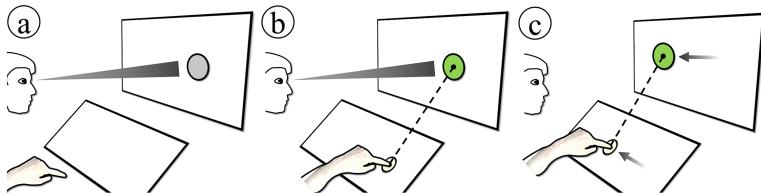
**Data Analysis.** For the statistical analysis of completion time and accuracy, we used a factorial repeated-measures ANOVA (Greenhouse-Geisser corrected if sphericity violated), and post hoc pairwise comparisons with Bonferroni corrections. For the analysis of Likert ratings, we used a Friedman test with a post hoc Wilcoxon Signed Rank test (Bonferroni corrected if necessary).

## 4 Experiment 1: Object Dragging

We first investigate an object dragging task. Dragging is commonly used on touchscreens, such as for moving files into a folder, or when positioning icons. In our experimental task, users move objects from a start to a destination location.

### 4.1 Compared Interaction Techniques

- **Touch:** Select the object with direct touch, and drag the object to the destination. The object starting and destination position are on the close, direct display.
- **Gaze+touch<sub>1:1</sub>:** Look at the object, touch down anywhere, indirectly drag object to the destination (Fig. 4). The task is performed on the remote screen, while touches are issued on the close display. The finger movement translates 1:1 to object movement on remote screen).
- **Gaze+touch<sub>dynamic</sub>:** Same procedure (Fig. 4), but the movement translates relatively using a dynamic control-display (CD) gain. This amplifies dragging i.e. it is more precise at slow and faster at fast finger movement (based on Windows XP/Vista pointer ballistics, similar to [5] ).



**Fig. 4.** Drag with Gaze+touch: look at the object (a), touch down (b), and drag (c).

## 4.2 Experimental Design

The object's size is set to  $3^\circ$  which is 31 mm on close ( $31_{mmc}$ ) and 49 mm on remote display ( $49_{mmr}$ ). It appears towards a corner of the screen, and the dragging direction is always toward the opposite diagonal corner. The target of the dragging was displayed as a grey circle (see Fig. 3b). The independent variables were:

- Dragging distance:  $17.5^\circ$  ( $185_{mmc}$ ,  $292_{mmr}$ ), and  $35^\circ$  ( $1210_{mmc}$ ,  $1307_{mmr}$ ) (between start and destination).
- Destination size: 110 % ( $34_{mmc}$ ,  $54_{mmr}$ ), 220 % ( $69_{mmc}$ ,  $109_{mmr}$ ), and 330 % ( $103_{mmc}$ ,  $164_{mmr}$ ) of the object size.

Overall, each user performs 90 trials: 3 techniques  $\times$  2 distances  $\times$  3 target sizes  $\times$  5 repetitions. All tasks were successfully completed by all users.

## 4.3 Results

- **Task Completion Time:** Task completion times are presented in Table 1, line 1–3. A main effect of technique ( $F(2,30) = 47.9$ ,  $p < .001$ ) showed that Touch is significantly faster than both Gaze+touch techniques ( $p < .001$ ), and Gaze+touch<sub>dynamic</sub> was faster than Gaze+touch<sub>1:1</sub> ( $p < .05$ ). Participants were slower with increasing distance ( $F(1,15) = 86.26$ ,  $p < .001$ ). An interaction effect between technique and distance ( $F(1.3,19.3) = 9.58$ ,  $p < .01$ ) showed Touch was the fastest technique across both distances ( $p < .01$ ). Further, Gaze+touch<sub>dynamic</sub> was faster than Gaze+touch<sub>1:1</sub> on  $35^\circ$  distance ( $p < .001$ ), which can be accounted to CD gain.
- **Accuracy:** Accuracy is the distance between the object's final position and center of the destination, normalized to degree of visual angle to allow direct comparison. Results are listed in Table 1, line 4–6. A main effect of technique ( $F(1.9,28.8) = 5.17$ ,  $p < .05$ ) showed that users were overall more accurate with Touch than both Gaze+touch techniques ( $p < .05$ ). Further, users were more accurate with shorter distances ( $F(1,15) = 23.63$ ,  $p < .001$ ) and larger target sizes ( $F(2,30) = 23.91$ ,  $p < .001$ , all pairs differ,  $p < .01$ ).

**User Feedback:** Participants' preferences were split during the dragging task, with 9 of the 16 users preferring Touch. The users' rationale for Touch was less mental demand ("my attention is not completely absorbed [with Touch]"), familiarity with this technique ("I have been using touchphones for a while, so it is a familiar technique"), ease, and speed of Touch ("it was easy and quick to perform the

**Table 1.** Quantitative results of the experiments ('green' denotes higher performance, 'light green' denotes higher performance than 'white', asterisks denote significance, 'time' measured in seconds, 'accuracy' measured in visual angle) (Color figure online).

Experiment	Line	Dependent Variable	Interaction Technique		
			Gaze+touch (1:1)	Gaze+touch (dynamic)	Touch
			Mean (SD)	Mean (SD)	Mean (SD)
<b>Dragging</b>	1	Time (17.5°)	3.12 (1.35)	2.9 (1.07)	2.11 (.93)*
	2	Time (35°)	4.24 (1.56)*	3.64 (1.56)*	2.57 (1.01)*
	3	Time (all)	3.68 (1.56)*	3.27 (1.39)*	2.34 (1)*
	4	Accuracy (17.5°)	0.44 (.55)	0.54 (.66)	0.38 (.37)
	5	Accuracy (35°)	0.64 (.76)	0.54 (.56)	0.46 (.58)
	6	Accuracy (all)	0.54 (.67)	0.54 (.61)	0.42 (.61)*
<b>Rotation (Varying Object Size)</b>	7	Skipped tasks	0%		19%
	8	Time (2°)	4.36 (1.59)*		6.23 (2.74)
	9	Time (4°)	4.25 (1.68)		4.24 (1.41)
	10	Time (8°)	4.43 (1.56)		4.04 (1.22)
	11	Time (all)	4.35 (1.6)*		4.8 (2.13)
	12	Accuracy (2°)	1.08 (.94)*		1.71 (1.43)
	13	Accuracy (4°)	0.81 (.75)		0.98 (.88)
	14	Accuracy (8°)	0.55 (.52)		0.59 (.52)
	15	Accuracy (all)	0.81 (.78)*		1.08 (1.1)
	16	Time (all)	5.79 (2.16)		5.81 (2.51)
	17	Accuracy (all)	0.74 (1.07)*		0.87 (.88)
	18	Skipped tasks	0%	0%	24.7%
	19	Time (2°)	3.64 (1.47)*	4.31 (1.79)	4.52 (1.82)
	20	Time (4°)	3.75 (1.39)	3.98 (1.67)	3.68 (1.5)
	21	Time (8°)	4.6 (1.76)	4.65 (1.94)	3.77 (1.4)*
<b>Scaling</b>	22	Time (all)	4 (1.6)	4.32 (1.85)	3.99 (1.62)
	23	Accuracy (2°)	0.072 (.05)	0.073 (.042)	0.105 (.074)
	25	Accuracy (4°)	0.074 (.059)	0.078 (.059)	0.098 (.07)
	26	Accuracy (8°)	0.082 (.064)	0.099 (.077)	0.102 (.086)
	27	Accuracy (all)	0.076 (.058)*	0.083 (.062)*	0.102 (.07)*

task"). Arguments in favour of Gaze+touch were no occlusion ("Touch had the problem that you obscured the object"), the speed of the eyes ("the eyes were faster than the fingers on the screen"), and less perceived effort ("[with Touch] it felt there was more effort because you saw the hand moving"). Gaze+touch<sub>dynamic</sub> was preferred over Gaze+touch<sub>1:1</sub> by 13 of the 16 participants due to less physical movement with the hands.

- **Questionnaire:** The questionnaire's result is presented in Table 2, line 1–6. Statistical tests showed that for the category of perceived ease ( $X^2(2,16) = 12.1$ ,  $p = .002$ ), Touch was perceived easier than the Gaze+touch<sub>1:1</sub> variant ( $Z = -2.7$ ,  $p < 0.05$ ). Regarding speed ( $X^2(2,16) = 12.1$ ,  $p = .002$ ), both Touch ( $Z = -2.76$ ,  $p < 0.05$ ) and Gaze+touch<sub>dynamic</sub> ( $Z = -2.49$ ,  $p < 0.05$ ) were perceived as faster than

**Table 2.** Qualitative results of the experiments (1 = strongly disagree, 5 = strongly agree, ‘green’ denotes better rating, asterisks denote significance) (Color figure online).

Experiment	Line	Dependent Variable	Interaction Technique		
			Gaze+touch absolute	Gaze+touch relative	Touch
			Mean (SD)	Mean (SD)	Mean (SD)
<b>Dragging</b>	1	Ease	3.81 (.83)	4.25 (.68)	4.75 (.45)*
	2	Speed	3.63 (.81)	4.25 (.68)*	4.44 (.73)*
	3	Accuracy	3.81 (.98)	4.06 (.57)	4.25 (.58)
	4	Learning	4.19 (.66)	4.06 (1.06)	4.88 (.34)*
	5	Eye tired.	3.63 (.81)	3.81 (1.05)	4.62 (.50)*
	6	Physical	3.81 (.91)	4.13 (.96)	3.69 (.79)
<b>Rotation (Varying Object Sizes)</b>	7	Ease		4.06 (.57)*	2.75 (1)
	8	Speed		3.94 (1)*	2.69 (.87)
	9	Accuracy		4.31 (.79)*	3.00 (.73)
	10	Learning		4.38 (.50)	4.13 (.72)
	11	Eye tired.		3.50 (.97)	4.25 (1)*
	12	Physical		3.50 (.97)	3.00 (.97)
<b>Rotation (Orientations)</b>	13	Ease		3.94 (.77)*	3.38 (.81)
	14	Speed		3.88 (.96)	3.50 (.63)
	15	Accuracy		4.00 (.82)	3.81 (.66)
	16	Learning		4.25 (.68)	4.06 (.93)
	17	Eye tired.		3.38 (1.31)	4.44 (.51)*
	18	Physical		3.50 (.97)*	2.75 (1)
<b>Scaling</b>	19	Ease	4.44 (.51)*	4.13 (.72)*	2.75 (.86)
	20	Speed	4.44 (.51)*	4.13 (.72)*	3.06 (.85)
	21	Accuracy	4.25 (.58)*	4.44 (.51)*	2.81 (.75)
	22	Learning	4.50 (.52)	4.38 (.72)	3.94 (.85)
	23	Eye tired.	3.19 (.98)	3.31 (1.01)	4.25 (.68)*
	24	Physical	3.69 (.95)	3.94 (.57)	3.25 (.93)

Gaze+touch<sub>1:1</sub>. Regarding learning effort ( $X^2(2,16) = 13.6$ ,  $p = .001$ ), users find Touch easier to learn than Gaze+touch<sub>dynamic</sub> ( $Z = -2.8$ ,  $p < 0.05$ ) and Gaze+touch<sub>1:1</sub> ( $Z = -2.8$ ,  $p < 0.05$ ).

For eye tiredness ( $X^2(2,16) = 14.4$ ,  $p = .001$ ), users find Touch less tiring than Gaze+touch<sub>dynamic</sub> ( $Z = -2.57$ ,  $p < 0.05$ ) and Gaze+touch<sub>1:1</sub> ( $Z = -2.86$ ,  $p < 0.05$ ).

#### 4.4 Discussion

Users were faster and more accurate with Touch, and Gaze+touch<sub>dynamic</sub> was faster than Gaze+touch<sub>1:1</sub>. User preferences did not show clear results, however they indicate that more users prefer using Touch. Further analysis of the questionnaire showed that

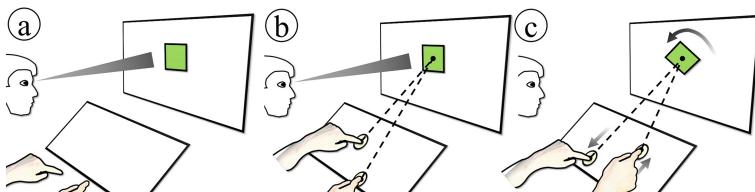
users perceive Touch to be beneficial across most categories (ease, speed, accuracy, etc.). Regarding both Gaze+touch techniques in comparison, users preferred Gaze+touch<sub>dynamic</sub> over Gaze+touch<sub>1:1</sub>, for requiring less hand movement.

## 5 Experiment 2: Object Rotation (Varying Object Sizes)

The previous study investigated a single-touch context; we now study multitouch interaction beginning with rotation of objects. We also investigate the fat-finger problem: we hypothesise that Touch is affected, but not Gaze+touch as fingers can be placed more freely. The task is a two-finger rotation tasks. Participants had to select an object, and then rotate it to a specific orientation. In each task, the object's size and rotation angle is varied.

### 5.1 Compared Interaction Techniques

- **Touch:** The user directly puts two fingers on an object, then rotates it to a target orientation. Again, touches and manipulation occur on the close and direct display. After users selected a target, users can also expand their fingers to manipulate the target more freely (as in all the following rotation and scaling experiments).
- **Gaze+touch:** The user looks at the target, indirectly touches down two fingers, then rotates these fingers to rotate the selected object (Fig. 5).



**Fig. 5.** Gaze+touch: look at the object (a), touch down two fingers (b), and rotate (c).

### 5.2 Experimental Design

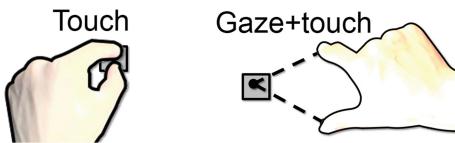
The target orientation was displayed underneath the object to visually indicate the rotation direction and angle (grey box, Fig. 3c). The object and the target both displayed a line, which the users had to match to finish the rotation. The independent variables were:

- Object size (visual angle): 1° (10<sub>mmc</sub>, 16<sub>mmr</sub>), 2° (20<sub>mmc</sub>, 33<sub>mmr</sub>), 4° (41<sub>mmc</sub>, 66<sub>mmr</sub>), and 8° (83<sub>mmc</sub>, 123<sub>mmr</sub>).
- Rotation angle: 10°, 50°, and 90°.

The smallest size of  $1^\circ$  is chosen as a realistic lower limit of the used eye tracker. If users struggled with the acquisition of this target with Touch (as  $1^\circ = 25$  px), they could skip the task. Overall, each user performs 120 trials: 2 techniques  $\times$  4 sizes  $\times$  3 rotation angles  $\times$  5 repetitions.

### 5.3 Results

- **Error:** All tasks were successfully completed with Gaze+touch, and users skipped 19 % of tasks with Touch. In particular, 76 % of tasks with a size of  $1^\circ$  were skipped, demonstrating the effect of the fat-finger problem. The error is illustrated in Fig. 6. For the following statistical analysis, we excluded all trials with  $1^\circ$  to ensure an equal numbers of conditions.



**Fig. 6.** Users skipped more tasks with Touch, as the fat-finger issue became apparent with small targets.

- **Task Completion Time:** Table 1 line 8–11 summarise task completion times. A main effect of technique ( $F(1,15) = 6.12$ ,  $p < .05$ ) showed that overall users were significantly faster with the Gaze+touch technique ( $p < .05$ ). A main effect of object size ( $F(1.3,19.3) = 40.77$ ,  $p < .001$ ) showed that object size affected task completion time: tasks with  $2^\circ$  sized objects were significantly slower than  $4^\circ$  and  $8^\circ$  ( $p < .001$ ), yet no difference was found between  $4^\circ$  and  $8^\circ$ . Similarly with rotation angle ( $F(2,30) = 17.23$ ,  $p < .001$ ), task completion time decreased at  $10^\circ$  ( $p < .01$ ), yet  $50^\circ$  and  $90^\circ$  did not show significant differences. An interaction between technique and size ( $F(1.5,21.9) = 35.68$ ,  $p < .001$ ), revealed that participants performed faster with Gaze+touch at the  $2^\circ$  tasks than with Touch ( $p < .001$ ).
- **Accuracy:** Accuracy is measured in degrees as the difference between the final angle of the object and that of the destination. The results are presented in Table 1, line 12–15. A main effect of technique ( $F(1,15) = 20.32$ ,  $p < .001$ ) showed that Gaze+touch was significantly more accurate than Touch ( $p < .001$ ). Accuracy increased with increasing object size ( $F(2,30) = 129.59$ ,  $p < .001$ , all pairs differ at  $p < .001$ ), but not significantly with rotation angle. An interaction between technique and size ( $F(2,30) = 12.47$ ,  $p < .001$ ) showed that for target size  $2^\circ$  Gaze+touch was significantly more accurate than Touch ( $p < .01$ ).
- **User Feedback:** 15 of 16 users preferred Gaze+touch. Most users' reasons were based around the fat-finger problem ("I wasn't constrained by finger size").
- **Questionnaire:** The questionnaire's result is shown in Table 2, line 7–12. The following statistical differences were revealed. For category ease ( $X^2(1,16) = 12$ ,

$p = .001$ ), Gaze+touch was perceived as easier than Touch. For category speed ( $\chi^2(1,16) = 13, p = .0$ ), Gaze+touch was perceived as faster than Touch. For category accuracy ( $\chi^2(1,16) = 16, p = .0$ ), Gaze+touch was perceived as more accurate than Touch. For category eye tiredness ( $\chi^2(1,16) = 7.4, p = .007$ ), Gaze+touch was perceived as more eye tiring than Touch.

## 5.4 Discussion

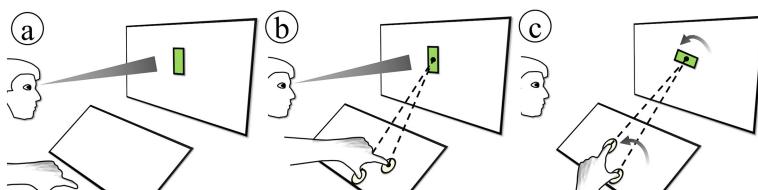
Users skipped more tasks when selecting small targets with Touch, a result that was expected considering the fat-finger issue. Gaze+touch did not show any errors across all target sizes, indicating that a gaze-selection is more accurate than a raw Touch-selection for two-touch gestures. Users were faster with Gaze+touch for small targets as they were easier to acquire (size  $\leq 2^\circ$ ). In case of accuracy, users performed more accurate than Touch in the rotation tasks with small targets, which can be accounted to the more relaxed placement of fingers. User feedback confirms these results, most users preferred Gaze+touch for being less constrained to finger size.

## 6 Experiment 3: Object Rotation (Varying Object Orientation)

The previous study investigated rotation of objects with varying object sizes. Another important factor for rotation is the initial orientation of targets which can affect the user's performance [8]. We therefore investigate this context in our third experiment, where users perform unimanual rotation gestures. In each task, the object's initial orientation and the rotation direction are varied.

### 6.1 Compared Interaction Techniques

- **Touch:** The user directly puts two fingers of the same hand on an object, then rotates it to a target rotation.
- **Gaze+touch:** The user looks at the target, indirectly touches down two fingers of the same hand, then rotates these fingers to rotate the selected object (Fig. 7).



**Fig. 7.** Rotate with Gaze+touch: look (a), touch down two fingers of one hand (b), and rotate (c).

## 6.2 Experimental Design

The target object is a rectangle with a width of 2° (20<sub>mmc</sub>, 33<sub>mmr</sub>) and a height of 10° (104<sub>mmc</sub>, 166<sub>mmr</sub>), and the angle of required rotation is fixed at 90°. We chose 2° as it allowed users to comfortably place their fingers on the object. The target orientation was indicated underneath the object with a rectangle and with an arrow to visually indicate the rotation direction and angle (Fig. 3). Independent variables were:

- Object starting orientation: 0°, 45°, 90°, and 135° relative to the screen's x-axis.
- Rotation direction: clockwise or anticlockwise.

Overall, users perform 90 trials: 2 techniques × 4 initial orientations × 2 directions × 5 repetitions. The participants completed all tasks with both techniques.

## 6.3 Results

- **Task Completion Time:** No significant effects were found for task completion times; users performed similarly fast with both techniques (Table 1, line 16).
- **Accuracy:** A significant main effect of technique for accuracy (again measured in angle difference,  $F(1,15) = 19.43$ ,  $p < .01$ ) revealed that Gaze+touch was more accurate than Touch (Table 1, line 17).
- **User Feedback:** 12 of 16 users preferred rotation with Gaze+touch, with the reason that Touch, with differently oriented objects, made some objects hard to acquire (“*You do not need to put the hand on the object*”).
- **Questionnaire:** Results are shown in Table 2, line 13-18. In category ease of use, users perceived Gaze+touch as easier than Touch ( $X^2(1,16) = 5.4$ ,  $p = .02$ ). For category eye tiredness ( $X^2(1,16) = 6.4$ ,  $p = .011$ ), Gaze+touch was perceived as more eye tiring than Touch. For category physical effort ( $X^2(1,16) = 6.4$ ,  $p = .011$ ), Gaze+touch was perceived as less physically demanding than Touch.

## 6.4 Discussion

Overall, users performed equally fast with both techniques, but more accurately with Gaze+touch. Contrasting the first rotation study (Experiment 2), no difference in speed is found because the target size remained fixed for all tasks in this experiment. However, users were more accurate with Gaze+touch, again to be accounted to the relaxed finger placement, as users indicated objects are difficult to acquire with Touch. The questionnaire showed that users perceived Gaze+touch as easier and as less physically demanding, although more eye tiring.

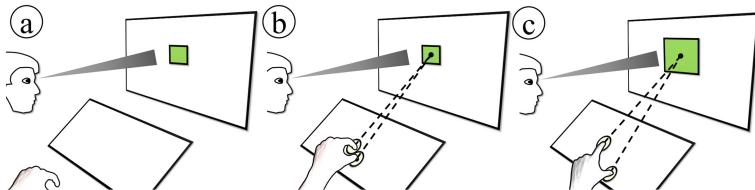
## 7 Experiment 4: Object Scaling

Our previous experiments investigated dragging and rotation tasks. Another prevalent gesture for multitouch is pinch-to-scale, often used to scale images, maps, or other objects. This experiment assesses users' performance of Gaze+touch in a scaling task.

Object size and scaling amplitude is varied. We let users choose between uni- or bi-manual interaction.

## 7.1 Compared Interaction Techniques

- **Touch:** The user directly puts two fingers on an object, then pinches them to fit an outline of a differently scaled target.
- **Gaze+touch<sub>1:1</sub>:** The user looks at the target, indirectly touches down two fingers, then pinches them to scale (Fig. 8). The scaling translates absolutely, thus finger movement is 1:1 applied to the object's scaling.
- **Gaze+touch<sub>dynamic</sub>:** Same operation (Fig. 8). However, the scaling translates with a different control-display gain, the distance between both fingers is relative to the width of the object.



**Fig. 8.** Gaze+touch scaling: look (a), touch down two fingers (b), and pinch (c).

## 7.2 Experimental Design

Users had to scale an object to a specific size which was visually indicated by a black rectangle (see Fig. 3). The independent variables were:

- Initial object size (visual angle): 1° (10<sub>mmc</sub>, 16<sub>mmr</sub>), 2° (20<sub>mmc</sub>, 33<sub>mmr</sub>), 4° (41<sub>mmc</sub>, 66<sub>mmr</sub>), and 8° (83<sub>mmc</sub>, 123<sub>mmr</sub>).
- Scaling factor: 50 %, 90 %, 120 %, and 200 % of the initial object's size.

Overall, a user did 240 trials: 3 techniques × 4 sizes × 4 scale factors × 5 repetitions.

## 7.3 Results

- **Error:** Participants completed all tasks with Gaze+touch, and skipped 24.7 % tasks with Touch (c.f. Figure 6). In particular, users skipped with Touch within 1° tasks 100 % of both downscaling conditions, 68.8 % upscaling-to-120 % tasks, and 56.3 % upscaling-to-200 % tasks; and within 2° tasks 29.7 % downscaling-to-50 % tasks, and 40.6 % downscaling-to-90 % tasks. Like in Experiment 2, we exclude skipped-task conditions from statistics for an equal number of trials between techniques.

- **Task Completion Time:** Task completion times results are summarised in Table 1, line 19-22. There was no significant difference for technique at task completion time. A significant main effect for object size ( $F(2,30) = 7.68, p < .01$ ) showed that users performed faster with  $4^\circ$  than  $8^\circ$  objects ( $p < .01$ ). A main effect of target size ( $F(1,15) = 34.98, p < .01$ ) revealed that users were faster with 120 % than 200 % upscaling ( $p < .05$ ). An interaction between technique and the object's size ( $F(2,33.8) = 7.6, p < .01$ ) showed that within  $2^\circ$  sized object tasks, users performed faster with Gaze+touch<sub>1:1</sub> than with Gaze+touch<sub>dynamic</sub> ( $p < .05$ ), and within  $8^\circ$  tasks, users were faster with Touch than with either Gaze+touch technique ( $p < .05$ ).
- **Accuracy:** Accuracy (the difference between radius of the object and radius of the target, normalized to degree of visual angle) showed a main effect of technique ( $F(2,30) = 9.8, p < .01$ ) with users more accurate with both Gaze+touch techniques than with Touch ( $p < .05$ , Table 1, line 23–27).
- **User Feedback:** All users preferred using one of the Gaze+touch techniques. The user's reason were based around the fat-finger problem ("I could not get both fingers within the box"), occlusion, and hand fatigue ("my arms do not get so tired because I can rest on the desk"). Between both Gaze+touch techniques, 8 users preferred Gaze+touch<sub>dynamic</sub>, and the other 8 users Gaze+touch<sub>1:1</sub>. Reasons were more precision with Gaze+touch<sub>dynamic</sub> ("I was more precise"), and speed for Gaze+touch<sub>1:1</sub> ("[Gaze+touch<sub>dynamic</sub>] is slower [than Gaze+touch<sub>1:1</sub>], because your fingers need to make a bigger movement").
- **Questionnaire:** Results are presented in Table 2, line 19–24. For category ease ( $X^2(2,16) = 26, p = .0001$ ), Gaze+touch<sub>dynamic</sub> ( $Z = -3.4, p < 0.05$ ) and Gaze+touch<sub>1:1</sub> ( $Z = -3.5, p < 0.05$ ) were perceived as easier than Touch. For category speed ( $X^2(2,16) = 19.4, p = .0001$ ), Gaze+touch<sub>dynamic</sub> ( $Z = -2.9, p < 0.05$ ) and Gaze+touch<sub>1:1</sub> ( $Z = -3.3, p < 0.05$ ) were perceived as faster than Touch. For category accuracy ( $X^2(2,16) = 24.6, p = .0001$ ), Gaze+touch<sub>dynamic</sub> ( $Z = -3.4, p < 0.05$ ) and Gaze+touch<sub>1:1</sub> ( $Z = -3.3, p < 0.05$ ) were perceived as more accurate than Touch. For category eye tiredness ( $X^2(2,16) = 16.8, p = .0001$ ), Gaze+touch<sub>dynamic</sub> ( $Z = -3, p < 0.05$ ) and Gaze+touch<sub>1:1</sub> ( $Z = -3.1, p < 0.05$ ) were perceived as more tiring than Touch.

## 7.4 Discussion

Similar to experiment rotation (size), users had difficulties acquiring small targets with Touch (fat-finger issue). Despite small targets, users performed similarly fast with Gaze+touch and Touch, while more accurate with Gaze+touch. Only for large targets, Touch was faster than Gaze+touch. Users' reasons were about the fat-finger problem (occlusion, fatigue). Considering both Gaze+touch variants, users can scale smaller objects faster with Gaze+touch<sub>1:1</sub>. User opinion was split, half of them preferred Gaze+touch<sub>1:1</sub> and the remaining users preferred Gaze+touch<sub>dynamic</sub>.

## 8 Overall Discussion

Our experiments investigated how Gaze+touch combines properties of direct (same gesture operation) and indirect touch (relaxed finger placement), which we follow up with a high level discussion on the following points:

**Advantages of Gaze+touch.** The relaxed finger placement can reduce effects of the fat-finger and occlusion issues. This was particularly beneficial for rotation and scaling tasks, where more than one finger is used for manipulation. Our experiments showed that with the design techniques, users performed similarly fast, and more accurately than with Touch. Manipulation of objects with a size of  $< 2^\circ$  was not feasible with Touch, however with Gaze+touch, users were able to manipulate the  $1^\circ$  objects (although completion time increased).

**Shortfalls of Gaze+touch.** Gaze+touch was slower and less accurate for dragging tasks. Based on observations and feedback, we believe this is accounted to two aspects. First, the ‘leave-before-click’ issue [17]: users can already look away from a target before they touch down, or touch down before they look at the target; both will void a Gaze+touch selection. Second, users could ‘lose’ a target during dragging, e.g. when the system wrongly detected a ‘touch up’ event or the user’s finger briefly hovered. With direct touch, the finger is on the target, thus the system will immediately receive a ‘touch down’ event and the dragging continues. However, with Gaze+touch, it is indirect touch, and to reselect users have to look to the target before they touch down. Both advantages and shortfalls correlated with the users’ feedback.

**Control-Display Gain.** The relaxed finger placement allowed us to experiment with dynamic CD gains (Gaze+touch<sub>dynamic</sub>). During dragging, this led to faster performance than a 1:1 mapping of Gaze+touch. It eliminated the need to clutch when the touch screen’s size did not suffice, confirming previous studies of cursor acceleration [5]. CD gain in scaling tasks did not use acceleration, but rather input relative to the fingers’ distance. Users were faster with the absolute scaling, but more precise with relative scaling.

**Observations.** We observed users exploiting the relaxed finger placement as a strategy to improve their performance: for long rotations, users had their fingers drawn together to manipulate faster, while for short rotations, users set them further apart to be more accurate. Although technically normal touch allows the same operation, it is restricted by a lower limit, as for small targets users have little room to adjust their fingers on it, and an upper limit, defined by the object’s size (although users can expand their fingers once they selected an object).

**Limitations.** Users were slower with Gaze+touch for dragging tasks. One method to overcome this could be using gaze for target dragging: after a touch selection, the target follows the user’s gaze ([22, 24, 25]). Future studies can investigate how this advanced technique would compare to standard touch dragging. In addition, our experiments compared Gaze+touch against the raw default of direct touch. This is apparently prone to issues such as the fat-finger problem or occlusion. These issues can be overcome, e.g. by indirect touch techniques ([1, 4]), and should be considered in future evaluation.

## 9 Conclusion

Gaze as a mediator can bring touch to remote displays, but at what cost and benefit? As a first step, we compared Gaze+touch to touch in four experiments across dragging, rotation, and scaling tasks. Our experiments provide detailed performance characteristics of both input modes, and indicate that while Gaze+touch is slower and less accurate for dragging tasks, users are equally fast and more accurate in rotation and scaling tasks. This can support the design of Gaze+touch interaction for remote displays, such as combined close and remote surfaces consistently controlled by multi-touch. In light that further evaluation beyond abstract tasks may be required to validate the real-world applicability, our experiments provide empirical groundwork in the exploration of how Gaze+touch sets itself apart from touch interaction.

## References

1. Albinsson, P.-A., Zhai, S.: High precision touch screen interaction. In: CHI 2003, pp. 105–112. ACM (2003)
2. Au, O.K.-C., Su, X., Lau, R.W.H.: LinearDragger: a linear selector for target acquisition on touch screens. In: CHI 2014, pp. 2607–2616. ACM (2014)
3. Benko, H., Izadi, S., Wilson, A.D., Cao, X., Rosenfeld, D., Hinckley, K.: Design and evaluation of interaction models for multi-touch mice. In: GI 2010, pp. 253–260. CIPS (2010)
4. Benko, H., Wilson, A.D., Baudisch, P.: Precise selection techniques for multi-touch screens. In: CHI 2006, pp. 1263–1272. ACM (2006)
5. Casiez, G., Vogel, D., Balakrishnan, R., Cockburn, A.: The impact control-display gain on user performance in pointing tasks. *Hum. Comput. Interact.* **23**, 215–250 (2008)
6. Forlines, C., Wigdor, D., Shen, C., Balakrishnan, R.: Direct-touch vs. mouse input for tabletop displays. In: CHI 2007, pp. 647–656. ACM (2007)
7. Gutwin, C., Cockburn, A., Scarr, J., Malacria, S., Olson, S.C.: Faster command selection on tablets with FastTap. In: CHI 2014, pp. 2617–2626. ACM (2014)
8. Hoggan, E., Williamson, J., Oulasvirta, A., Nacenta, M., Kristensson, P.O., Lehtiö, A.: Multi-touch rotation gestures: performance and ergonomics. In: CHI 2013, pp. 3047–3050. ACM (2013)
9. Jacob, R.J.K.: What you look at is what you get: eye movement-based interaction techniques. In: CHI 1990, pp. 11–18. ACM (1990)
10. Jacob, R.J.K.: Eye movement-based human-computer interaction techniques toward non-command interfaces. *Adv Hum. Comput. Interact.* **4**, 151–190 (1993). Ablex Publishing
11. Kin, K., Agrawala, M., DeRose, T.: Determining the benefits of direct-touch, bimanual, and multifinger input on a multitouch workstation. In: GI 2009, pp. 119–124. CIPS (2009)
12. Malik, S., Laszlo, J.: Visual touchpad: a two-handed gestural input device. In: ICMI 2004, pp. 289–296. ACM (2004)
13. Moscovich, T., Hughes, J.F.: Indirect mappings of multi-touch input using one and two hands. In: CHI 2008, pp. 1275–1284. ACM (2008)
14. Pfeuffer, K., Alexander, J., Chong, M.K., Gellersen, H.: Gaze-touch: combining gaze with multi-touch for interaction on the same surface. In: UIST 2014, pp. 509–518. ACM (2014)

15. Pfeuffer, K., Vidal, M., Turner, J., Bulling, A., Gellersen, H.: Pursuit calibration: making gaze calibration less tedious and more flexible. In: *UIST 2013* (2013)
16. Potter, R.L., Weldon, L.J., Shneiderman, B.: Improving the accuracy of touch screens: an experimental evaluation of three strategies. In: *CHI 1988*, pp. 27–32. ACM (1988)
17. Salvucci, D.D., Anderson, J.R.: Intelligent gaze-added interfaces. In: *CHI 2000*, pp. 273–280. ACM, New York (2000)
18. Schmidt, D., Block, F., Gellersen, H.: A comparison direct and indirect multi-touch input for large surfaces. In: *INTERACT 2009*, pp. 582–594. Springer (2009)
19. Sibert, L.E., Jacob, R.J.K.: Evaluation of eye gaze interaction. In: *CHI 2000*, pp. 281–288. ACM (2000)
20. Špakov, O., Isokoski, P., Majaranta, P.: Look and lean: accurate head-assisted eye pointing. In: *ETRA 2014*, pp. 35–42. ACM, New York (2014)
21. Stellmach, S., Dachselt, R.: Look & touch: gaze-supported target acquisition. In: *CHI 2012*, pp. 2981–2990. ACM (2012)
22. Stellmach, S., Dachselt, R.: Still looking: investigating seamless gaze-supported selection, positioning, and manipulation distant targets. In: *CHI 2013*, pp. 285–294. ACM (2013)
23. Turner, J., Alexander, J., Bulling, A., Gellersen, H.: Gaze + RST: integrating gaze and multitouch for remote rotate-scale-translate tasks. In: *CHI 2015*. ACM (2015, to appear)
24. Turner, J., Alexander, J., Bulling, A., Schmidt, D., Gellersen, H.: Eye pull, eye push: moving objects between large screens and personal devices with gaze & touch. In: *INTERACT 2013*, pp. 170–186. Springer (2013)
25. Turner, J., Bulling, A., Alexander, J., Gellersen, H.: Cross-device gaze-supported point-to-point content transfer. In: *ETRA 2014*, pp. 19–26. ACM (2014)
26. Voelker, S., Wacharamanotham, C., Borchers, J.: An evaluation state switching methods for indirect touch systems. In: *CHI 2013*, pp. 745–754. ACM (2013)
27. Ware, C., Mikaelian, H.H.: An evaluation of an eye tracker as a device for computer input. In: *CHI 1987*, pp. 183–188. ACM (1987)
28. Wigdor, D., Benko, H., Pella, J., Lombardo, J., Williams, S.: Rock & rails: extending multi-touch interactions with shape gestures to enable precise spatial manipulations. In: *CHI 2011*, pp. 1581–1590. ACM (2011)
29. Zhai, S., Morimoto, C., Ihde, S.: Manual and gaze input cascaded (MAGIC) pointing. In: *CHI 1999*, pp. 246–253. ACM (1999)

# Gestu-Wan - An Intelligible Mid-Air Gesture Guidance System for Walk-up-and-Use Displays

Gustavo Rovelo<sup>(✉)</sup>, Donald Degraen, Davy Vanacken, Kris Luyten,  
and Karin Coninx

Hasselt University - tUL - iMinds, Expertise Centre for Digital Media,  
Wetenschapspark 2, 3590 Diepenbeek, Belgium

{gustavo.rovelorui, donald.degraen, davy.vanacken, kris.luyten,  
karin.coninx}@uhasselt.be

**Abstract.** We present *Gestu-Wan*, an intelligible gesture guidance system designed to support mid-air gesture-based interaction for walk-up-and-use displays. Although gesture-based interfaces have become more prevalent, there is currently very little uniformity with regard to gesture sets and the way gestures can be executed. This leads to confusion, bad user experiences and users who rather avoid than engage in interaction using mid-air gesturing. Our approach improves the visibility of gesture-based interfaces and facilitates execution of mid-air gestures without prior training. We compare *Gestu-Wan* with a static gesture guide, which shows that it can help users with both performing complex gestures as well as understanding how the gesture recognizer works.

**Keywords:** Gesture guide · Mid-air gestures · Walk-up-and-use

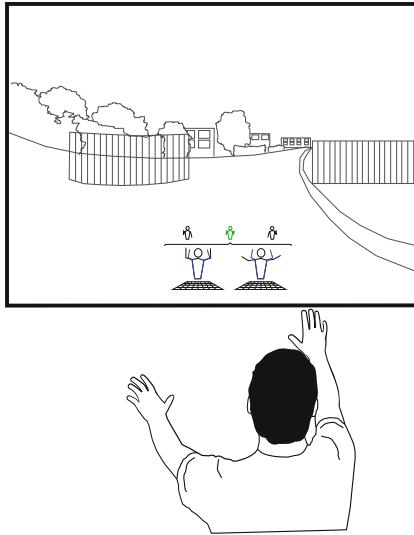
## 1 Introduction

Walk-up-and-use displays that provide mid-air gesture-based interfaces often struggle with informing their users of the possible gestures that can be executed and how to perform these gestures [20]. In this paper, we propose *Gestu-Wan*, an intelligible mid-air gesture guidance system for walk-up-and-use displays. Figure 1 shows *Gestu-Wan*, visualizing a mid-air gesture set to control an omnidirectional video<sup>1</sup>.

2D gesture-based interfaces have to cope with similar issues, and several solutions exist to improve the visibility of 2D gestures. For mid-air gestures, however, there are no comprehensive solutions that increase the visibility of the available gestures, provide guidance on how to execute those gestures and are tailored for usage without prior training.

In particular walk-up-and-use displays require users to be able to use the system without prior training [20]. This implies that users should feel comfortable interacting with a system they never encountered before, and of which they

<sup>1</sup> A 360° video recording, e.g. <http://www.youtube.com/watch?v=ClAuhgFQpLo>.



**Fig. 1.** *Gestu-Wan* is an intelligible gesture guidance system designed to support mid-air gesture-based interaction for walk-up-and-use displays.

cannot predict the behaviour. Ideally, potential users are informed of how to use such a display when they approach it, while also being enticed to use the system. This is a challenging task when interaction is accomplished exclusively by means of mid-air gestures. As a result, interaction designers usually limit both the number of gestures as well as their complexity, in order to improve the so-called “guessability” of gestures [22]. The goal of this work is similar to StrikeAPose [20]: revealing mid-air gestures for walk-up-and-use contexts, such as public displays, but where StrikeAPose focuses on discoverability of the initial gesture, we focus on guided exploration of the full gesture set and on informing the user about how the gesture recognizer works. In this paper, we focus on a single-user setting. Exploring multi-user scenarios is an important next step, as it requires an in-depth analysis of aspects such as floor control and mutual awareness.

*Gestu-Wan* shows users how to perform mid-air gestures of various degrees of complexity without prior explanation or training. It uses a minimalistic visualization of the upper body skeleton and a visual structure that shows users how to execute mid-air gestures before and during the actual gesture execution. The visualization has been created using an informed design method, surveying users on what elements are required to visualize a mid-air gesture. We incrementally added additional visual information until we reached a visual representation that offers just enough information to comfortably perform the gesture while minimizing visual clutter. We strived for a minimalist design for several reasons: (1) to be able to present a structured representation of how to perform various gestures on a limited screen size, (2) to avoid demanding too much attention

of the users or distracting users that have no need for the gesture guide, (3) to allow for integration with the content displayed and (4) to maximize the data-ink ratio. The latter reason is known to lead to clear and concise visualizations, and focuses on showing the core information that needs to be conveyed to the user [17].

Our solution provides an intelligible approach for revealing mid-air gestures. First, *Gestu-Wan* makes both the initial gesture as well as all available options visible to the user in exchange for a limited amount of screen estate. Secondly, *Gestu-Wan* provides feedforward similar to OctoPocus [3], but instead of presenting the full path of the gesture, it only shows the possible next steps (or poses) that can be performed and all the possible outcomes, given the current posture. Thirdly, *Gestu-Wan* provides real-time feedback while the user is performing interactions and reveals how a gesture is executed in all three dimensions. Finally, *Gestu-Wan* improves the user's awareness of how the gesture recognizer works, so the user will be able to quickly get used to the required accuracy of the gestures. We believe this to be of great value, since the way mid-air gesture recognizers work is hard to unravel for an end-user (i.e. required granularity and speed to perform the gestures), which leads to mismatches between the gesture performed and the gesture recognized.

## 2 Related Work

Without proper guidance, gesture-based interfaces have a steep learning curve, as most of the time gestures are difficult to discover and learn [9, 12, 22]. Several solutions have been proposed to improve the visibility and usability of 2D and mid-air gesture-based interfaces. For 2D gesture-based interfaces, such as mouse-based and multi-touch gestures, dynamic and real-time guidance systems have been proposed to support gesture execution (e.g. [3, 5, 7, 18]). For mid-air gestures, however, there are less guidance systems available, especially for walk-up-and-use interaction. Some notable examples that we will discuss in this section are LightGuide [15], YouMove [2], and StrikeAPose [20].

Even in very early work on 2D gesture-based interfaces, extra clues such as *crib-notes* and *contextual animations* were used to expose the available gestures and how they can be performed [9]. Marking menus [9] expand on pie menus by offering the user the possibility to draw a stroke towards the desired item in order to activate the associated command, thus integrating 2D gestures with menu selection. The user learns how to perform the gesture by following the menu structure while performing the gesture, and can eventually activate menu items blindly by performing the associated gesture. Hierarchical marking menus [10] increase the total number of menu options by presenting the user with several submenus. The subdivision of (complex) gestures is also a must for mid-air gestures, as it helps users to “find their way” through the different substeps to successfully execute a gesture.

Bau and Mackay focus on feedforward and feedback to facilitate learning and execution of complex gesture sets [3]. Their gesture guide, OctoPocus, is a

dynamic guide for single-stroke 2D gestures. Bau and Mackay are among the first to explicitly discuss feedforward as a separate concept for designing a successful gesture guide, an approach we also adopt in our work. Arpège [8] provides finger by finger feedforward and feedback to help users learn and execute static multi-finger postures on multi-touch surfaces. TouchGhosts [18] is a gesture guide for 2D multi-touch gestures that demonstrates available gestures through virtual animated hands that interact with the actual user interface to show the effects of the gestures. ShadowGuides [7] visualize the user's current hand posture combined with the available postures and completion paths to finish a gesture.

Anderson and Bischof [1] compared the learnability and (motor) performance of different types of 2D gesture guides, and found that approaches with higher levels of guidance exhibit high performance benefits while the guide is being used. They also designed a gesture guide that supports the transition from novice to expert users, because most existing system insufficiently support that transition. This work provides additional grounding for a high-level dynamic gesture guide as a prerequisite to achieve a usable walk-up-and-use interface with mid-air gestures.

Nintendo Wii, Microsoft Xbox Kinect and PlayStation Move games make extensive use of mid-air gesturing to control games. These systems provide the user with written and graphical representations of the movements that are needed to play the games. Most games also incorporate a “training stage”, an initial phase of the game that is specifically designed to teach users how to interact. For general gestures, Microsoft also provides a set of instruction movies<sup>2</sup>. Mid-air gestures in games are, however, typically very simple and do not require accurate movements. In a walk-up-and-use scenario, similar instructions can be displayed on the screen when users approach it, but that is only practical in case of a small and simple gesture set.

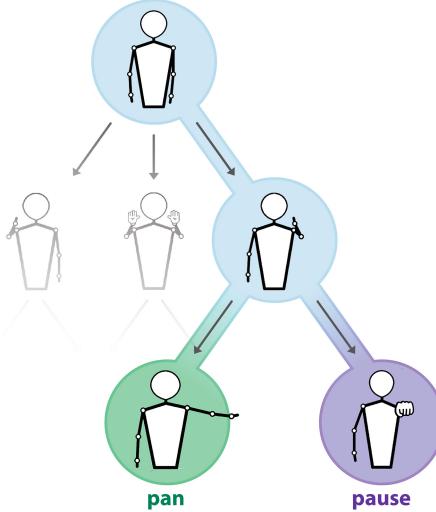
In Augmented Reality (AR) applications, mid-air gestures are also very common and wide-ranging, as shown by the very extensive user-defined gesture set for AR of Piumsomboon et al. [13]. White et al. [21] present graphical AR representations of potential actions and their consequences, which specifically target discovery, learning, and completion of tangible AR gestures.

Walter et al. [20] investigated how to reveal an initial mid-air *teapot* gesture on public displays. Permanently displaying a textual instruction and a static contour of someone performing this gesture was found to be the most effective strategy, but the gesture is very straightforward, as users only need to touch their hip. *LightGuide* [15] incorporates feedback and feedforward by projecting graphical hints on the user's hand to guide mid-air movements. Users need to focus on their hand, however, which makes *LightGuide* suitable for exercise or physical therapy, but less appropriate when users are interacting with an application that requires visual attention. Furthermore, Sodhi et al. did not explore guidance of two hands simultaneously.

*YouMove* [2] is a system for full-body movement and posture training. It presents the user with a skeleton representation that needs to be mimicked by

---

<sup>2</sup> <http://support.xbox.com/en-US/xbox-360/kinect/body-controller>.



**Fig. 2.** Gestures are subdivided in sequences of postures, which are structured hierarchically. A user follows a path in the gesture tree, performing the subsequent postures, until a leaf node is reached and the action associated with that gesture is executed. Note that this is only an illustrative visualization of the underlying gesture tree; it is not a depiction of *Gestu-Wan*.

the user, and uses an interactive mirror that shows the user in real-time how her current posture matches the target posture. *YouMove* is specifically built for training, however, and does not facilitate the discovery of the available gestures. Similar to *LightGuide*, users only have to focus on the guidance itself, since there is no underlying application that requires visual attention. This is an important difference, because it is challenging to clearly show users how to perform mid-air gestures without distracting them too much from the main purpose of the application.

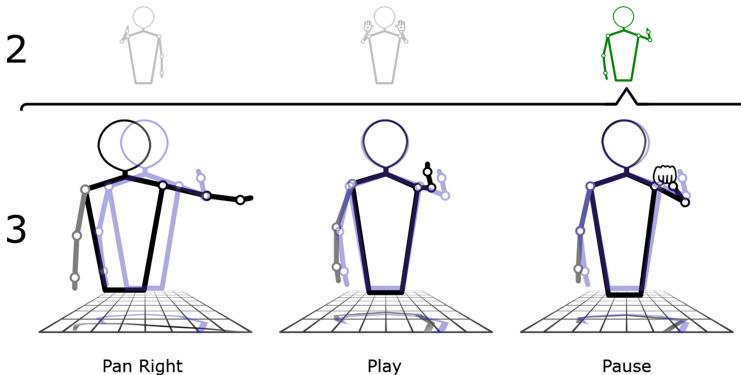
In summary, *Gestu-Wan* is the first guidance system for walk-up-and-use scenarios that provides feedforward and feedback for mid-air gestures. We refer the reader to the work of Delamare et al. [6] for a complete discussion of the design space for gesture guidance systems.

### 3 An Intelligible Gesture Guide

In this section, we discuss how *Gestu-Wan* assists users in discovering and performing mid-air gestures, and its specific properties that make it suitable for walk-up-and-use situations.

#### 3.1 Structure and Visual Appearance

*Gestu-Wan* uses a set of intermediate postures to generate the guidance visualization. We define this dataset in a pre-processing phase by manually subdividing

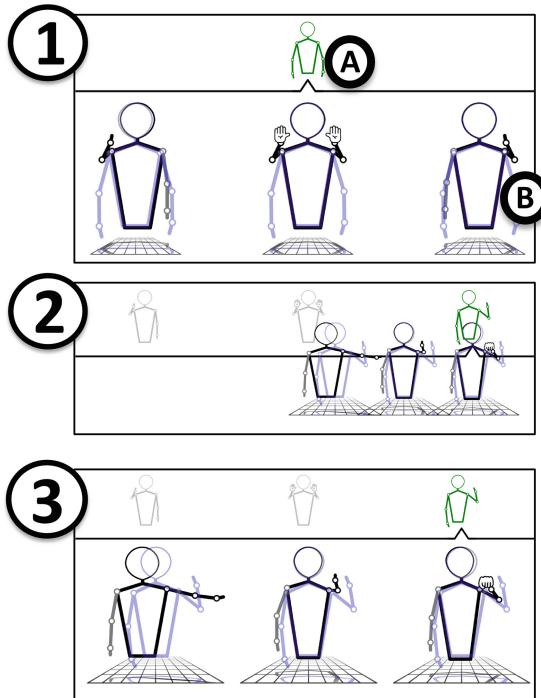


**Fig. 3.** *Gestu-Wan* showing a set of possible mid-air gestures. Each gesture is subdivided in a number of steps, thus a sequence of postures that need to be matched. The user can see the previous step (the green skeleton at the top), the next possible steps (the black skeletons), and her current posture (the blue overlay skeleton). The numbers at the side represent the steps: the user already executed two steps and is now in the third step (Color figure online).

each mid-air gesture in a number of steps, a sequence of postures that needs to be matched. By decomposing the gestures, they can be *structured hierarchically* in a tree, as shown in Fig. 2. A user follows a path in the gesture tree, performing the subsequent steps that are shown in the nodes, until a leaf node is reached and the action associated with that gesture is executed.

Segmenting gestures can be cumbersome, as not every type of gesture can be decomposed in a straightforward manner. Consider, for example, a continuous circular movement, or varying the zoom level according to the distance between the hands of the user. Splitting up such a continuous gesture or representing the parametric input causes an exponential growth of the tree or requires adding an extra level of detail (e.g. arrows, labels) to the representation. Although *Gestu-Wan* is able to support both approaches, we did not further investigate this feature in the context of the presented work. The automatic segmentation of gestures into series of reproducible postures is also outside the scope of this paper. Our focus is on the representation and structuring of mid-air gestures, and on showing what effect the execution of a gesture has.

Instead of showing the entire gesture tree to the user, *Gestu-Wan* only shows the previous posture that the user correctly matched and the next possible postures, as seen in Fig. 3. Each of the next possible postures is also labelled with text that describes which action(s) can be reached through that posture. The current posture of the user is displayed in blue and is continuously updated, while the possible target postures are shown in black. Once the current posture of the user matches with one of the target postures, either the next possible steps of the gesture are shown, or, in case of a leaf node, the action corresponding to the gesture is executed. Note that we only add an open or closed hand to the



**Fig. 4.** Step (1): (A) shows the previously matched posture, while (B) is the target posture. Step (2): When the target posture is matched, it moves up and the tree expands to show the next steps. Step (3): All the next possible postures are shown and the user can match one of them to continue with the gesture.

skeleton when that particular hand posture is required; when the hand is not visualized, the recognizer does not take into account the hand posture.

If the current posture of the user no longer matches the requirements from the previous steps, the color changes from blue to red and the user has a few moments to correct her posture. If not corrected in time, the gesture tree resets to the first step.

To avoid unintentional execution of gestures when a new user approaches the system, *Gestu-Wan* first shows an initial gesture that the user has to perform in order for the recognizer to start interpreting the postures in the gesture tree. The importance of learning about such an initial gesture for walk-up-and-use displays has been shown by Walter et al. (the *teapot* gesture) [20].

### 3.2 Animations

*Gestu-Wan* provides two types of animations that contribute to the recognizability of how to proceed to finish a gesture: the *overlay skeleton* and *hierarchy traversal* animations. The overlay skeleton, depicted in blue in Fig. 4, shows the

current posture of the user on top of the target posture, shown in black. The current posture is tracked in real-time, so any movements of the user are immediately reflected in the overlay skeleton. This dynamic aspect is important for users to identify themselves with the overlay skeleton and to quickly identify what posture to take. Performing the next step of a gesture is thus reduced to trying to match the overlay posture with the target posture.

The traversal through the gesture hierarchy, also depicted in Fig. 4, is animated: each time a target posture is matched, the gesture tree shifts upwards. The top shows the users where they came from (i.e. which posture they matched in the previous step) and the bottom shows what the next possible steps are. This strengthens the idea that the gesture set can be navigated as a hierarchy, which makes it easier for users to situate their current posture within the whole gesture tree.

### 3.3 Feedback and Feedforward

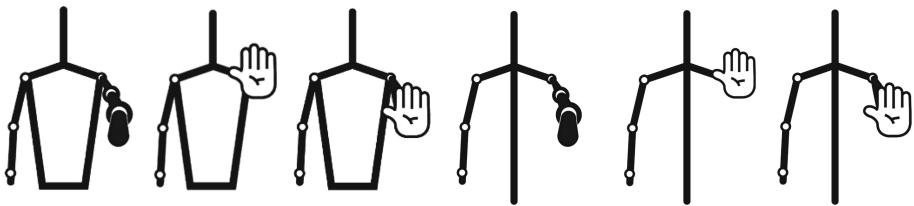
The core aspect to create an intelligible gesture guide is providing both feedback on the user’s actions and the system’s behaviour, as well as feedforward to show what effect the execution of a gesture has. While a vast amount of work exists on this in the area of 2D gestures (e.g. [1, 3, 10, 11]), only little can be found on using a combination of feedforward and feedback for mid-air gestures.

Similar to the feedforward provided by OctoPocus for 2D gestures [3], *Gestu-Wan* presents *continuous* and *sequential feedforward* while executing a gesture. It shows what the possible next steps are when reaching a particular posture, thus what postures can be performed next (*sequential*) while performing the posture (*continuous*). Besides the continuous and sequential feedforward, *Gestu-Wan* also includes functional affordances. This is simply done by using labels to indicate the possible actions, thereby revealing the effect one can expect (e.g. “zoom in” or “pan left”). For a full description of feedforward, we refer to Bau and Mackay [3] and Vermeulen et al. [19].

While feedforward tells the user what to do and what she can expect when she does this, feedback tells her more about how the system tracks her while performing gestures. The most obvious feedback is the overlay skeleton that tracks the user continuously and shows the tracked posture in real-time. Once a posture is matched, it turns green and *Gestu-Wan* shifts the hierarchy up by means of an animation. When the user’s current posture no longer matches the requirements from the previous steps, the overlay skeleton turns red until the user corrects her posture or the gesture tree resets.

## 4 Design Rationale

Since we target walk-up-and-use settings, the system should provide guidance in such a way that occasional and first time users can easily perform mid-air gestures while immediately engaging with the system. It is, however, very challenging to clearly show users how to perform mid-air gestures without overwhelming



**Fig. 5.** Several of the designs that respondents of the online survey had to evaluate according to clarity and visual attractiveness.

them with details or distracting them too much from the main purpose of the application. In an early design phase, in which five users tried a first version of *Gestu-Wan*, we identified three categories of challenges: (1) providing appropriate depth cues (occlusions, shadows, perspective, etc.), (2) showing which body parts contribute to a gesture, and (3) visualizing gestures that include hand postures.

Depth cues are needed to show 3D movements and postures. When a rich and wide-ranging set of mid-air gestures is required, the gestures inevitably become more complex and require more movements in all three dimensions, which makes depth cues indispensable. However, a broad range of depth cues exist, and we want to find the minimum amount of details required to clearly convey depth. Showing which body parts are involved in a gesture is important to avoid that users try to match the entire body, and not just the parts that matter. For instance, a gesture might simply require the user to raise and close her left hand. However, if the gesture guide does not explicitly visualize the right hand as “inactive”, the user might also try to precisely match the position and posture of her right hand, which makes the gesture unnecessarily hard and uncomfortable to perform.

With these challenges in mind, we performed an online survey to guide the design of the guidance visualization. In this survey, 50 respondents rated different skeleton representations to find out the best combination of depth cues, body representation, level of detail for the hand postures. Participants were asked to rate each of these designs on a scale according to clarity and visual attractiveness. Figure 5 shows an example of three design alternatives.

The respondents preferred a complete trapezoid representation of the upper body, including both arms, the torso and the head, using a different shade of color to represent an inactive arm. Respondents argued that using a more complete representation of the body increases the “guessability” of the gestures, and that the torso and its rotation could be easily perceived. Drawing only a stick figure was confusing, as participants expressed that the position of the arms with respect to the body was less clear. The respondents also highlighted the importance of occlusions to perceive depth (i.e. the torso being partially hidden by the representation of the arms, hands and joints). A perspective projection of the skeleton was clearer than using an orthogonal projection, and using shadows

(i.e. a projection of the torso, arms and joints) combined with a rectangular grid drawn with a perspective view was the clearest way to provide depth cues.

In addition to the results from the online survey, we considered the relevant guidelines from the set suggested by Anderson et al. for YouMove [2]. *Gestu-Wan leverages domain knowledge* on the behaviour of the gesture recognizer. The hierarchical structure shows how gestures are composed of subsequent postures, and the animations show how the recognizer progresses to the next steps while performing series of postures. This not only improves the visibility of the recognizer’s behaviour, but also makes the interaction more intelligible [4], since the system shows the user how it perceives and processes the input. The user will notice early why some movements and postures work or do not work, and are thus able to efficiently interact with the system much faster.

The overlay skeleton immediately shows the user that she is being tracked as soon as she enters the field of view, which improves the discoverability of the interactive nature of the system. As a result, users quickly connect with the system, thereby *motivating the user* to engage in further interactions. *Gestu-Wan* conveys the necessary information to smoothly start interacting without prior training, while the visualization minimizes visual clutter, *keeping the presentation simple*. There is no need for any prior experience with mid-air gesturing, and gestures do not need to be remembered, requiring a *low cognitive load*.

## 5 User Study

The main goal of the user study is to evaluate to what extent the properties of *Gestu-Wan* enable walk-up-and-use interaction. In particular, we want to evaluate if the visual structure and the continuous feedforward and feedback increase discoverability and learnability of mid-air gestures, and, at the same time, increase awareness of the recognizer’s behaviour.

We performed a controlled lab experiment, in which we asked volunteers to perform a set of tasks in an Omni-Directional Video (ODV) player. The content of the ODV is a first-person video walk-through of university premises (some streets, buildings and parking lots), providing an experience similar to Google Street View, but using video instead of a set of still images. The ODV player supports the following actions: *play*, *pause*, *restart*, *pan left and right*, *zoom in* and *zoom out*.

The gesture set used in our study is based on the user-defined gesture set of Rovelo et al. [14], which specifically targets ODV interaction. According to their findings, the gestures are reasonably easy to learn and in some cases also easily discoverable. Since all the gestures in that set are fairly simple (i.e. consisting of three “steps”), we replaced the pause and restart gestures with two more complex gestures of five steps. This allows us to test *Gestu-Wan* with a broader range of gestures and helps us to evaluate the depth dimension of the hierarchical representation more thoroughly. The added complexity of those gestures also prevents accidental activation of the “disruptive” action of restarting the ODV.

### 5.1 Baseline Considerations

*Gestu-Wan* is, to our knowledge, the first walk-up-and-use mid-air gesture guidance system, thus comparing it with another similar system is not possible. A few systems that support mid-air gestures, such as YouMove [2] and StrikeAPose [20], are available in literature and are discussed earlier in this paper. Although we highlighted some significant differences with these systems, the findings that are presented in those papers proved to be very useful as a basis for informing the design of *Gestu-Wan*. To evaluate *Gestu-Wan*, we decided to compare it against a static printed graphical gesture guide (see Fig. 6(A)). The paper version gives participants a complete overview of all gestures and allows them to quickly skim what they need to do without the time constraints that would be imposed by using, for example, video guides. We chose a static representation because it is a very low-threshold and informative way of representing gestures (e.g. always visible, easy to skim and read, no side effects of animations).

Alternatively, *Gestu-Wan* could be compared against a textual explanation of the gestures or a set of videos or animations. Both alternatives would require a lot of time to be processed by the participants and would not represent the whole gesture set in a visible, structured way. The printed graphical guide and *Gestu-Wan* do provide such an overview: the former shows the structure of the whole set at once, while the latter reveals it incrementally, while gestures are being performed. Furthermore, textual explanations and especially videos and animations do not offer accessible feedforward while performing gestures. The graphical printed guide does offer this in a sense, because the user can easily scan ahead in the graphical representation to see what is next and what it will lead to.

These considerations led us to choose the printed graphical gesture guide as a basis for comparison. For walk-up-and-use situations, this representation offers the most advantages over the other alternatives.

### 5.2 Methodology

We used a between-subjects design to compare the performance of participants who used the dynamic guide, *Gestu-Wan*, with participants who used static gesture representations printed on paper. The printed representations are similar to the drawings that are used by Nintendo Wii or Microsoft Kinect games, and clearly show the different steps that a user needs to perform to complete each gesture, as shown in Fig. 6.

Every participant started the study with an exploration phase, in which the participant was free to try and explore the system. After the participant performed the initial gesture, she could control the ODV player by performing any of the gestures that the (dynamic or static) guide presents. During this initial phase, we explained neither the gesture guide, nor the ODV player, in order to evaluate the effectiveness of the gesture guidance for walk-up-and-use interaction. We interrupted this initial phase after the participant performed each gesture at least once. If a participant kept repeating certain gestures or if

she did not explore every available gesture, she would be reminded that the next phase of the experiment would commence as soon as each gesture was executed.

Next, one of the observers explained the purpose of the study and the tasks they needed to perform. We reset the ODV to its original starting point and asked participants to perform each action three more times. The first two times, participants were asked to execute every action one by one, e.g. when asked to zoom out, the participant had to find the necessary steps to perform that particular action. The third time, participants were asked to accomplish a more complex task: “walk until building three is visible, then zoom in on its address sign and read it to us”. This task required several actions, such as starting the video, pausing it when the building is found, panning to focus on the area that contains the address sign and zooming in on the sign. By giving participants a task that requires them to engage with the ODV content, we created a situation that resembles using the system in a real-life setting, which helped us to investigate the user experience in the post-usage questionnaire.

After performing all the tasks, participants answered a short questionnaire regarding the gesture guide and their experience with the interaction with the ODV. They also exchanged their thoughts with one of the observers during an informal interview, for instance on the gestures that caused them problems.

The system logged the time every participant spent on each step of the gestures and the number of times that participants activated the wrong action. Two observers were present during the study and took notes. Afterwards, the observers analyzed video recordings of the sessions to annotate participants’ questions, remarks and frustrating moments during their interaction with the system. The observers also used this analysis phase to validate the annotations made during the study.

### 5.3 Participants

We invited 26 participants (15 men and 11 women). We balanced both conditions with respect to gender and professional background: 15 participants had a background in computer science, six in bio-science, three in economy and administration, one in statistics and one in graphical design.

All participants were familiar with 2D gesture-based interfaces, as they all own an smartphone. Experience with mid-air gestures was limited, as only two participants of the baseline group reported that they play Microsoft Xbox Kinect games (one plays less than one hour a week, the other between one and five hours). Six participants reported they play video games using the Nintendo Wii: one of the *Gestu-Wan* group and two of the baseline group play less than one hour, while two of the former group and one of the latter group play between one and five hours a week.

Given our focus on walk-up-and-use displays, we opted for a wide diversity of participants and did not perform a visual-spatial awareness test beforehand. The high number of participants with a background in computer science is because of the environment in which the study took place. There were, however, also no participants who regularly use mid-air gestures among the computer scientists.

## 5.4 Apparatus

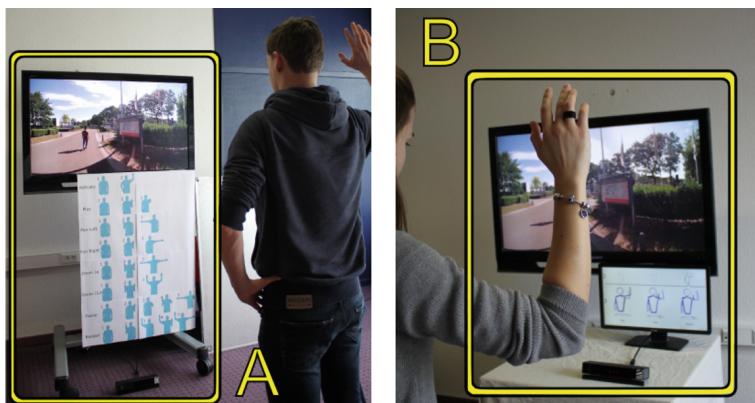
We used a 43" display to show the ODV, mounted on a pedestal to place its center more or less orthogonal to the participants line of sight, as shown in Fig. 6. *Gestu-Wan* was shown on a separate 21.5" screen, placed at the lower right corner of the ODV screen. The printed guide was on a A0 sheet, located below the ODV screen. The size of the representation of each gesture was the same in both conditions (approximately 12 by 7 cm).

Participants were standing at a fixed position, two meters from the display. To track the participants' movements, we used a second-generation Microsoft Kinect. We also recorded all the sessions with two video cameras. One camera was used to record the gestures and the participants' comments while they were interacting. The other camera recorded the ODV and gesture guide.

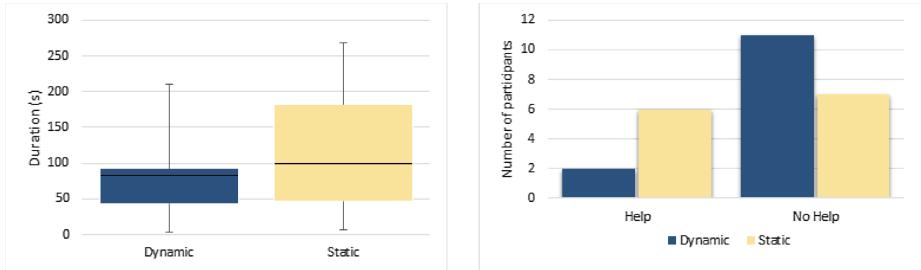
## 5.5 Results and Discussion

*Phase 1: Free Exploration Phase.* The free exploration phase offers us the possibility to compare how both types of guidance help participants to discover the interaction possibilities without any prior knowledge of the system. We extracted the time to activate the system from the system's logs, by calculating the elapsed time between the moment when the participant was instructed to start and the moment when she successfully performed the initial gesture. This gives us some insights into how long it takes users to get acquainted with the system.

Figure 7(a) shows the distribution of elapsed time for both conditions. Welch's t-test shows no significant effect of the guidance on the time to activate the system ( $t(24) = -1.102$ ,  $p = 0.141$ ). With the dynamic guide, the average time was  $81.37 \pm 63.65$  seconds, while it was  $116.34 \pm 95.08$  seconds with the static guide. Note that this elapsed time not only encompasses the time that was required to execute the initial gesture, but rather the whole initial phase to discover and get



**Fig. 6.** On the left, the setup that was used for the study with the printed gesture guide (A). On the right, the setup with *Gestu-Wan* (B).



(a) Distribution of the time that was required to activate the system during the free exploration phase of the user study.

(b) Number of participants who required assistance from one of the observers to activate the system during the free exploration phase of the user study.

**Fig. 7.** Results for the exploration phase.

acquainted with the system, since this was the first time that participants saw the setup. Furthermore, there was no sense of urgency to start interacting with the ODV. Participants who used *Gestu-Wan* typically took their time to first explore the behaviour of the skeleton (e.g. moving, leaning forward, waving). This type of “playful” behaviour was also reported by Tomitsch et al. [16]. We did not observe this playful behaviour in participants who used the static guide. Some of these participants were confused about the status of the system (e.g. asking “Is the system working already?”) or asked the observers for feedback and instructions on what to do.

A number of participants could not activate the system without assistance from one of the observers: two participants in case of the dynamic condition and six in case of the static condition (Fig. 7(b)). We provided extra feedback to them about the cause (e.g. not lifting the arm sufficiently high), so they could continue with the study.

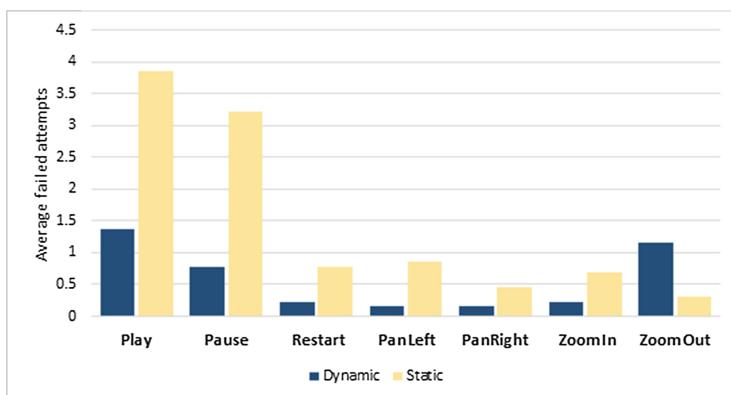
If we look at the gestures performed after the initial gesture, one participant of each group was not able to perform the “play” gesture, and one participant who used the static guide could not perform the “restart” gesture. The most difficult gesture during exploration seemed to be “pause”: while only one participant using the dynamic guide failed to perform it, 11 of the 13 participants who used the static guide did not succeed. Fisher’s exact test reveals that the number of participants who completed the “pause” gesture differs significantly between conditions ( $p < 0.001$ , 95 % CI[5.60, 5743.07], odd ratio = 82.68). Since we redefined the “pause” gesture from the gesture set of Rovelo et al. [14] and made it much more complex to execute, this effect is in line with our expectations.

*Phase 2: Individual Gesture Execution.* In this phase, participants were asked to execute every action one by one. We analyzed the video recordings and logs to count the number of failed attempts before they correctly performed a gesture. We marked an attempt as failed when a participant triggered the wrong action, or performed the sequence of steps, but one of the steps was not recognized by the system and the action was thus not triggered.

The analysis reveals a trend that indicates that the dynamic guide helps participants to be more accurate when performing gestures. Figure 8 shows an overview of the medium number of failed attempts per condition for each gesture. The order in the figure is also the order in which the participants had to execute the gestures. Only one participant could not perform the “pause” gesture with the dynamic guide. With the static guide, on the other hand, one participant could not perform “play”, 10 participants could not perform “pause”, and one participant could not perform “restart”. The considerably higher average number of failed attempts for the “zoom out” gesture in the dynamic condition is caused by one of the participants who failed to perform the gesture over 10 times in a row. This participant executed all other gestures fine, and we have no explanation for the sudden change. This distorts the results for the “zoom out” gesture, although when not including this outlier, the static condition still outperforms the dynamic condition slightly (mean number of attempts dynamic: 0.4, static: 0.3).

There is a statistically significant difference in the number of failed attempts for the “pause” gesture ( $t(17.17) = -3.767$ ,  $p = 0.001$ , 95 % CI[-3.84, -1.08]). This confirms our expectation that for more complex gestures (e.g. composed of more than three steps and requiring movements in the depth plane), the dynamic guide outperforms the static guide. For simple gestures, however, the performance of both guides does not differ statistically. Fisher’s exact test reveals that the number of participants who completed the “pause” gesture differs significantly between conditions ( $p < 0.001$ , 95 % CI[3.34, 2534.63], odd ratio = 41.35).

The analysis did not reveal any overall statistically significant difference regarding the time that participants required to perform every action. This might be due to the fact that following the guide in the dynamic condition costs time, but also because of the simplicity of most gestures.



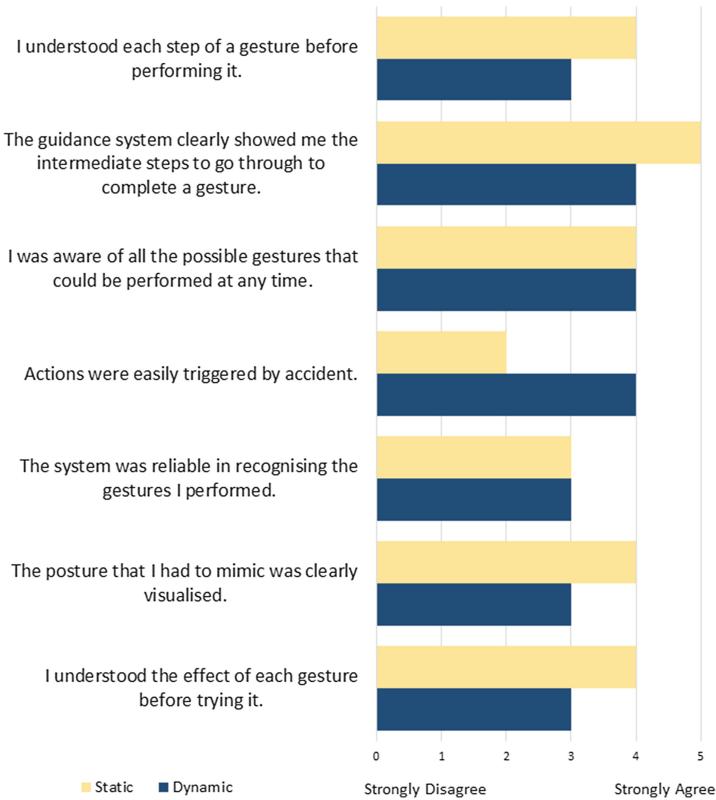
**Fig. 8.** Distribution of failed attempts to perform the seven gestures during the second phase of the user study. Complex gestures benefit from the dynamic condition, while there is no significant benefit for simple gestures.

*Phase 3: Navigate and Search.* Participants were asked to perform a navigation and search task, giving them the opportunity to evaluate the experience during a realistic task. They were asked to look for a specific location in the ODV, and then focus and zoom in on a specific point of interest. Participants of both groups rated the intrusiveness of the guides as neutral. However, our observations and the comments of four participants during the informal interviews show that participants were almost constantly observing either the dynamic or static guide while executing a gesture. Given the position of both the screen (dynamic) and paper (static), as shown in Fig. 6, this is not surprising. The out-of-context visualization interferes with watching the ODV content, since it requires shifting focus. We envision the dynamic gesture guide being integrated in the actual content, as depicted in Fig. 1, which should make it less disruptive.

*Post-study Questionnaire.* Our post-study questionnaire, of which an overview of the results is shown in Fig. 9, revealed some interesting and unexpected results. In the dynamic condition, users were more aware of the gestures that were executed. When they performed a gesture by accident, they were able to identify what happened because the gesture guide's feedback. A Mann-Whitney U-test shows a significant positive effect of the dynamic guide on awareness of gestures being accidentally performed ( $U = 132$ ,  $p = 0.013$ ). In both conditions, participants triggered actions by accident at about the same rate, but the median rating with regard to the question on triggering actions by accident is four for the participants who used the dynamic guide ("agree" on the five-point Likert scale), and two for the participants who used the static guide ("disagree" on the Likert scale).

The post-study questionnaire also shows that the dynamic guide provided adequate insight in the behaviour of the system, although it was rated somewhat lower than the static guide. This seems odd, since in the dynamic condition, the user is guided through the gestures step-by-step. However, the static guide shows an ordered overview of all postures required to perform a gesture, informing the user about the whole gesture at once. Moreover, since we used a between-subject design, we are unsure how users would rate the two conditions with respect to each other when being exposed to both guides. Being able to see all the steps at once also resulted in a higher subjective appreciation in case of the static condition. This indicates that a dynamic gesture guide might need to incorporate more extensive feedforward to provide a full overview of the gestures.

The static guide, however, takes up a lot more space than the dynamic guide, since all the steps of each gesture need to be presented at once. Even with the limited number of gestures used in our study, integration of the static guide with the display itself (e.g. overlayed on the content) is not possible. The gesture set that was used in the study only contained eight gestures, but the AR gesture set presented by Piumsomboon et al. [13] is, for instance, much more extensive. *Gestu-Wan* can handle such an extensive set more easily than a static guide, on the condition that we structure the gestures hierarchically. Furthermore, a larger gesture set typically requires users to perform gestures more accurately in order for the recognizer to distinguish potentially similar gestures. *Gestu-Wan* helps users to deal with a recognizer that requires accuracy.



**Fig. 9.** Summary of the post-experiment questionnaire.

## 6 Conclusion

We presented *Gestu-Wan*, a dynamic mid-air gesture guide specifically designed for walk-up-and-use displays. The contributions are that *Gestu-Wan* (1) makes the available gestures visible to users, (2) facilitates the execution of mid-air gestures, and (3) turns the gesture recognizer into an intelligible system. The intelligibility results in users gaining a better understanding *while* performing gestures about what to do next and about how the gesture recognizer works. Given the walk-up-and-use context, we assume that there is no upfront training and that users want to access the system immediately.

Throughout the design and evaluation of *Gestu-Wan*, we perceived the difficulties of supporting mid-air gestures. Nowadays, typical gestures are fairly simple and can easily be performed with a static guide. Such a static guide, however, requires a lot of space. We also noticed that a dynamic gesture guide not necessarily leads to a significantly improved performance. Difficulties in perceiving depth and the amount of attention that a dynamic guide requires might be as demanding as trying to reproduce gestures presented on a static medium.

The guidance provided by *Gestu-Wan* did, however, outperform the static guide for complex gestures.

We believe that the further progression of mid-air gesture-based interfaces entails challenges to present information to the user about which gestures are available and how to execute them, in a limited though easily accessible space. *Gestu-Wan* is the first system that provides a functional solution for these challenges.

**Acknowledgments.** The iMinds ICON AIVIE project, with project support from IWT, is co-funded by iMinds, a research institute founded by the Flemish Government. We thank the participants of our study.

## References

1. Anderson, F., Bischof, W.F.: Learning and performance with gesture guides. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2013, pp. 1109–1118. ACM (2013)
2. Anderson, F., Grossman, T., Matejka, J., Fitzmaurice, G.: Youmove: enhancing movement training with an augmented reality mirror. In: Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology, pp. 311–320. ACM (2013)
3. Bau, O., Mackay, W.E.: Octopocus: a dynamic guide for learning gesture-based command sets. In: Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology, UIST 2008, pp. 37–46 (2008)
4. Bellotti, V., Edwards, W.K.: Intelligibility and accountability: human considerations in context-aware systems. *Hum. Comput. Interact.* **16**(2–4), 193–212 (2001)
5. Bennett, M., McCarthy, K., O’Modhrain, S., Smyth, B.: SimpleFlow: enhancing gestural interaction with gesture prediction, abbreviation and autocompletion. In: Campos, P., Graham, N., Jorge, J., Nunes, N., Palanque, P., Winckler, M. (eds.) INTERACT 2011, Part I. LNCS, vol. 6946, pp. 591–608. Springer, Heidelberg (2011)
6. Delamare, W., Coutrix, C., Nigay, L.: Designing guiding systems for gesture-based interaction. In: Proceedings of the 7th ACM SIGCHI Symposium on Engineering Interactive Computing Systems (to appear), EICS 2015. ACM (2015)
7. Freeman, D., Benko, H., Morris, M.R., Wigdor, D.: Shadowguides: visualizations for in-situ learning of multi-touch and whole-hand gestures. In: Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces, pp. 165–172 (2009)
8. Ghomi, E., Huot, S., Bau, O., Beaudouin-Lafon, M., Mackay, W.E.: Arpège: learning multitouch chord gestures vocabularies. In: Proceedings of the 2013 ACM International Conference on Interactive Tabletops and Surfaces, ITS 2013, pp. 209–218. ACM (2013)
9. Kurtenbach, G., Moran, T.P., Buxton, W.: Contextual animation of gestural commands. *Comput. Graph. Forum* **13**(5), 305–314 (1994)
10. Kurtenbach, G., Buxton, W.: The limits of expert performance using hierarchic marking menus. In: Proceedings of the INTERACT 1993 and CHI 1993 Conference on Human Factors in Computing Systems, CHI 1993, pp. 482–487 (1993)

11. Long, Jr., A.C., Landay, J.A., Rowe, L.A.: Implications for a gesture design tool. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 1999, pp. 40–47 (1999)
12. Norman, D.A., Nielsen, J.: Gestural interfaces: a step backward in usability. *Interactions* **17**(5), 46–49 (2010)
13. Piumsomboon, T., Clark, A., Billinghurst, M., Cockburn, A.: User-defined gestures for augmented reality. In: Kotzé, P., Marsden, G., Lindgaard, G., Wesson, J., Winckler, M. (eds.) INTERACT 2013, Part II. LNCS, vol. 8118, pp. 282–299. Springer, Heidelberg (2013)
14. Rovelo Ruiz, G.A., Vanacken, D., Luyten, K., Abad, F., Camahort, E.: Multi-viewer gesture-based interaction for omni-directional video. In: Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems, pp. 4077–4086 (2014)
15. Sodhi, R., Benko, H., Wilson, A.: Lightguide: projected visualizations for hand movement guidance. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 179–188 (2012)
16. Tomitsch, M., Ackad, C., Dawson, O., Hespanhol, L., Kay, J.: Who cares about the content? an analysis of playful behaviour at a public display. In: Proceedings of The International Symposium on Pervasive Displays, PerDis 2014, pp. 160:160–160:165 (2014)
17. Tufte, E.R.: *The Visual Display of Quantitative Information*. Graphic Press, Cheshire (1986)
18. Vanacken, D., Demeure, A., Luyten, K., Coninx, K.: Ghosts in the interface: meta-user interface visualizations as guides for multi-touch interaction. In: Tabletop 2008: Third IEEE International Workshop on Tabletops and Interactive Surfaces, pp. 81–84 (2008)
19. Vermeulen, J., Luyten, K., van den Hoven, E., Coninx, K.: Crossing the bridge over norman’s gulf of execution: revealing feedforward’s true identity. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2013, pp. 1931–1940 (2013)
20. Walter, R., Bailly, G., Müller, J.: Strikeapose: revealing mid-air gestures on public displays. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2013, pp. 841–850 (2013)
21. White, S., Lister, L., Feiner, S.: Visual hints for tangible gestures in augmented reality. In: Proceedings of the 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality, pp. 1–4 (2007)
22. Wobbrock, J.O., Morris, M.R., Wilson, A.D.: User-defined gestures for surface computing. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2009, pp. 1083–1092. ACM (2009)

# Natural Interaction with Video Environments Using Gestures and a Mirror Image Avatar

Christian Kray, Dennis Wilhelm, Thore Fechner<sup>(✉)</sup>, and Morin Ostkmap

Institute for Geoinformatics, Westfälische Wilhelms-Universität Münster,  
Münster, Germany

{c.kray,d.wilhelm,t.fechner,m.ostkmap}@uni-muenster.de

**Abstract.** Video environments are a promising option for a variety of applications such as training, gaming, entertainment, remote collaboration, or user studies. Being able to interact with these environments enables further applications and extends existing application scenarios. In this paper, we propose a novel interaction technique that combines natural gestures with mirror images of the user to allow for immersive interaction with video environments. The technique enables movement inside the 3D space depicted by the video as well as the placement and manipulation of virtual objects within the 3D space. We describe a potential application scenario, where interactive public displays are placed inside a scene by one user and then experienced by another user. We also briefly report on a user study evaluating the gesture set we defined for controlling movement within the video.

**Keywords:** Gestural interaction · Mirror image · Avatar · Video

## 1 Motivation

There are many scenarios, where a realistic visual simulation of a real environment is needed: training, gaming, entertainment, or remote collaboration are some examples. In addition, the design and evaluation of ideas and (context-dependent) systems can strongly benefit from such simulations. The design process of ubiquitous systems such as public display networks can also benefit from realistic simulations of the intended deployment area: instead of implementing a fully functional system or deploying screens in the real world, designers can create mock-ups using simulations and thus gather feedback from users early on in the development process. Immersive video environments are one option to convincingly simulate the real world. While they are easy to create and very realistic, they do not include semantic or geometric information. Movement in the depicted 3D space and interaction with objects shown in the footage is thus not realized easily.

In this paper, we propose an approach to overcome these issues. The approach also enables the injection of new content into the video footage and the subsequent experiencing of this content. Our system combines gestural interaction with a mirror image of the user that serves as an avatar within the video

environment. It thereby enables the intuitive selection of 3D locations shown in video environments as well as the placement of virtual objects inside the 3D space depicted by the video footage.

## 2 Related Work

There are different approaches to realize visually convincing simulations. One is to use virtual environments (VEs), i.e., computer generated scenarios based on a detailed 3D model. Another approach is to use photographs or video footage to generate an immersive experience. Synthetic 3D models allow for fine grained details and interaction, while the actual modeling requires a lot of work. While parts of the process can be automated, e.g., using 3D scanners [6], the overall effort is still considerable. Conversely, photographs or video footage provide a realistic (audio-)visual simulation and can be captured quite easily and effortlessly but interaction with the shown objects is limited.

Different means of interaction are available to move within the simulation, e.g., treadmills, and to manipulate objects shown in the simulation (such as gloves or voice commands). Gestures can also be used for this purpose. Särkelä et al. [9] analyzed different gestures and notions to let users navigate in a VE. Vogel and Balakrishnan [12] define a design space for freehand pointing and clicking interaction. Nancel et al. [8] investigated how to implement mid-air pan-and-zoom gestures on wall-sized displays. Benko and Wilson [3] analyzed multi-point mid-air gestures for omnidirectional immersive environments.

A key issue in creating immersive experiences is the lack of haptic feedback. Vogel and Balakrishnan suggest to compensate for this by using additional visual and auditory cues [12]. In general, humans perceive visual stimuli more pronounced than auditory or tactile ones [10]. One possible approach to provide additional visual feedback is to employ a mirror metaphor. Mirrors are commonly used, and most users are thus familiar with how to operate them. This metaphor has been used in a variety of contexts [1, 5, 7, 10, 11]. Uses include raising awareness of health issues, motivating behavior change, or simply using a mirror image to focus attention during interaction.

Similar to the approach presented by Ahn et al. [1], the approach proposed in this paper uses the user's mirror image as a video avatar. The avatar can be used to navigate within the VE and to manipulate virtual objects. In contrast to previous work, our system does not merely substitute a cursor with a mirror image. Our approach is focussed on providing an immersive experience by using intuitive gestures in combination with the video avatar. It supports both the creation of augmented scenes, where virtual objects are inserted into video footage and the exploration of such scenes. Gotardo and Price [4] aimed at a similar workflow and developed a system that is comparable to the one presented here. However, their approach is based on a much more complex hardware setup. Moreover, the user interface designed by Gotardo and Price is based on a heads-up-display (HUD) rather than on gestures to let users select actions (e.g., selecting or scaling objects). A purely gestural interface may be experienced as a

more natural way to interact with an immersive video environment (IVE), both by designers and participants. In addition, the system proposed here allows for a quick and easy creation of scenes from video footage and does not require custom and expensive hardware.

### 3 Approach

The main goals of our approach were to enable intuitive interaction with video environments and to immerse people watching video footage as much as possible into the real world scene being shown. Our aim was thus to enable users to perform various actions in the simulated environment while providing them with a strong feeling of presence. The basic idea underlying our proposed interaction technique is to create a realtime mirror image of the user and overlay it over the video footage. Using a simple, layer-based depth model and a small set of gestures, users can place their mirror avatar inside the depicted real-world scene and interact with the environment via the avatar (e.g., to place virtual objects). In the following paragraphs, we describe all essential components of the approach in detail and provide an overview of our prototypical implementation.

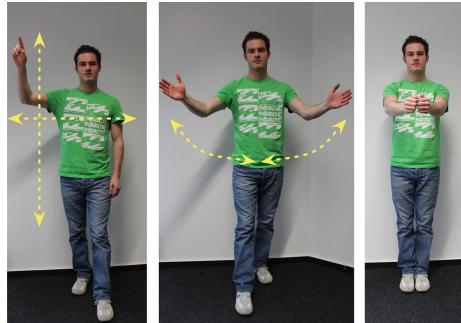
#### 3.1 Mirror Image Avatar

A live mirror image of the user constitutes the focal point of interaction, see Fig. 2(a)–(d). The background behind the user, which is also captured by the camera pointed at the user, is eliminated in real-time, e.g., using standard difference image or chroma-key techniques. The cut-out mirror image serves as an avatar or proxy for the user: its location in 3D space defines where interaction can take place.

In particular, the avatar defines the depth layer that is currently selected. Since both the user and the system's depth camera know how tall the user is, the actual size of the avatar as depicted on screen defines its depth position inside the video footage. The size of the avatar is used as a depth cue to inform the user about the layers and objects that can be selected. For example, by placing the avatar on the third layer three, the user can interact with objects located on that layer and can inject virtual objects on that layer.

#### 3.2 Gestures

**Avatar Control.** In order to move the avatar within the 3D space defined by the video footage and the layer model, users can perform a number of gestures. When one foot is placed in front of the other, users can move their avatar along the X-, Y-, and Z-axis. To control the movement in X- and Y-direction, users use a one-handed gesture. By extending one arm in one of the cardinal directions, a user can specify in which direction the avatar should move, see Fig. 1 (left). For example, extending the arm to the left moves the avatar in that direction.

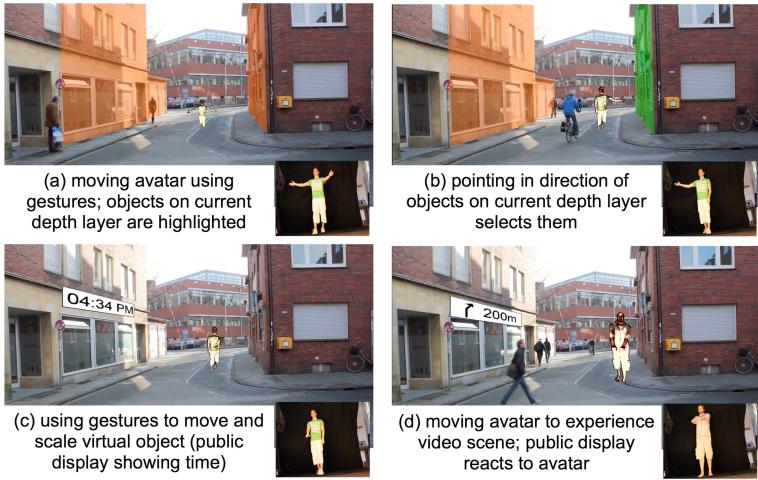


**Fig. 1.** Key gestures used to control the avatar, interact with the video scene, and to manipulate virtual objects: moving gestures (left), scaling gesture (middle), switching gesture (right).

Movement continues while the arm is extended. Movement stops when users bring their arm close to their bodies, or when they perform a different gesture.

A two-handed gesture controls movement along the Z-axis (depth). Putting both hands closely together in front of the body shrinks the avatar. This corresponds to moving it deeper into the image, i.e., it increases its distance from the camera. By spreading both arms, users can increase the size of the avatar and thus decrease its distance to the camera, see Fig. 1 (middle). The scaling stops when users either return their arms to a relaxed position, or when they perform another gesture. The size of the avatar determines which depth layer is selected: the depth information specified for each layer and the actual height of the user enables the system to compute the best match, i.e., to find the layer which corresponds best to the current size of the avatar.

In order to evaluate these gestures, we carried out a comparison study that contrasted it with an alternative set of gestures, where movement was controlled by walking in place while orienting one's body in the target direction. Twenty participants (ten male, ten female, average age 22.35 years, SD: 1.927) were recruited via word of mouth and other means from around the university. They were asked to navigate several immersive video scenes in the IVE with the help of the mirror image avatar using two different sets of movement gestures. All participants had to use both methods, but the order of exposure was randomized. Due to space constraints we cannot report on the entirety of the results here but can only summarise the key findings. Participants were largely successful in navigating the avatar to the target locations using either method. We frequently observed people stopping when their avatar reached a street and avoiding “collisions with cars,” which indicates a high degree of immersion. The gesture set depicted in Fig. 1 was rated more favorably (SUS score of 74.25, SD: 12.76) than the comparison set (SUS score of 64.63, SD: 23.13). 70 % of the participants also preferred the static gesture shown in Fig. 1 overall.



**Fig. 2.** Example scenario: (a) moving and scaling the avatar, (b) selecting scene element, (c) placing virtual object, (d) experiencing virtual object (e.g., public display).

**Object Manipulation.** In addition to moving the mirror image avatar, we defined a set of gestures to select objects depicted in the video footage (e.g., buildings or signs), to inject virtual objects (e.g., public displays, new buildings, or audio sources) as well to move and scale those objects. In addition, gestures to select content to inject into the scene or other content-related activities can be defined. For example, in our prototypical implementation we included gestures to select an item to inject from a list of options.

In order to select an object in the video footage, users first need to place their avatar on the corresponding layer. Users then select an object by simply pointing in the direction of the object. The gestures to place an object in 3D space are the same ones as those used for moving the mirror image avatar. Users can switch between moving their avatar and moving objects by putting their hand together, extending their arms in front of them and then maintaining this pose for a short time, see Fig. 1 (right). Visual feedback, i.e., a progress bar, indicates the switch from one set of actions to another.

**Experiencing Augmented Video Scenes.** Once video footage has been augmented with a number of virtual objects, people can experience the new “scenario” in the following way: by moving their mirror image avatar through the 3D space defined by the video footage and the layer model (as described above), they can interact with the objects augmenting the video scene. The layer model provides means to, for example, measure the distance to a public display and realize proxemic interaction [2]. The next section describes an example scenario, which demonstrates how users can experience augmented video scenes.

## 4 Example Scenario

We created a prototypical implementation of our approach using predominately web-based technologies and two cameras (a webcam and a depth camera), which runs in real-time inside a standard browser on a desktop PC. In order to demonstrate the use of our approach, we propose the following example scenario, see Fig. 2. Designers want to create a public display system that consists of a series of screens distributed throughout the city. These screens are meant to react to passersby and provide them with personalized information. Instead of installing real displays at the planned deployment sites, short video clips of those sites can be recorded and be used to prototype the system.

The first step is to construct the layer model: for each scene, the designers need to define a number of layers that are located at different distances (depth levels) from the plane defined by the camera. In addition, they can mark up a number of objects shown in the video footage and link them to a specific layer. Layers and objects define how people can later interact with the video footage. Once layers (and optionally objects) are specified, people can interact with the footage. Designers can now use the gestures shown in Fig. 1 to move their avatar around the video scene. As they move their avatar, previously defined objects are highlighted, see Fig. 2(a)—indicating, which objects can be “reached” from the current position of the avatar. When designers have positioned their avatar at the desired location, they can point at objects, which are associated with the layer the avatar is located on, see Fig. 2(b). Additionally, the designers can inject virtual objects into the footage. In the example scenario, the designers insert, move, and scale a mockup of a public display, see Fig. 2(c). The virtual object is associated with the layer on which the avatar is located.

The resulting augmented video footage can then be explored by other people, e.g. stakeholders or the people who commissioned the public display system. Figure 2(d) shows a stakeholder who is exploring the scene that the designers created previously. Using the gestures shown in Fig. 1, the stakeholder moves their avatar around the video scene. The public display reacts to the avatar, and when it’s within its activation area, the display content changes and shows directions targeted at the user. The stakeholder can thus experience the design in a realistic way and explore, for example, whether the activation area of a public display or its content are fit for the intended deployment location.

## 5 Discussion

The initial prototypical implementation suffers from a number of limitations. It currently only supports a small set of objects that can be placed inside the video scene, and at the moment, their orientation in space cannot be changed. In addition, only one user can interact with the system at the same time. Furthermore, movement of the avatar is not restricted so that users can place it in physically impossible positions (e.g., floating above ground), which could break immersion. Finally, both the avatar and virtual objects are simply overlaid over the video

footage: moving objects such as cars that intersect with these simply disappear behind them regardless of where they are supposed to be in the 3D space defined by the video. This is another aspect that can negatively affect immersion.

Most of these limitations can be addressed by improving the current implementation. A more generic import mechanism for assets to be injected could use a different set of gestures (or a mobile device) to allow for the use of arbitrary virtual objects. A more sophisticated and robust gesture recognition system would allow for rotation gestures and enable multi-user interaction. A more sophisticated layer model could also specify permissible locations of the avatar on each layer to prevent avatars from being moved to physically impossible locations. Realizing physically correct occlusions involving the avatar would require a deeper analysis of the video and a more sophisticated spatial model.

Finally, several limitations relate to the gestures we used. While the gestures we defined were learned quickly by the participants and positively received in our user study, further studies are required to identify the most immersive/intuitive set of gestures. So far, we only assessed a subset of all the gestures, i.e., movement control. In addition, we did not test whether the use of devices, e.g., mobile phones, to carry out certain actions, such as injecting virtual objects, would be more immersive/intuitive, neither on their own nor in combination with gestures. Further studies on these aspects are desirable as well.

Generally speaking, mirror image avatars could also be used with photographs or true 3D environments, i.e., virtual worlds. We chose to use video footage as we expected the real-time motion of the avatar to blend in more naturally with the movement naturally occurring on video footage, and would thus create a strong sense of presence and immersion. Using true 3D environments would allow for correct occlusions but constructing realistic virtual worlds requires a lot of effort. Compared to a desktop scenario, where a user would place objects and experience augmented video scenes, we argue that the gesture-based approach combined with a large screen provides a more realistic and immersive experience. Initial informal feedback from people seeing the system in action as well as observations from the initial user study on the movement gestures seem to confirm this, but we intend to carry out a series of user studies to investigate these aspects in more detail. Clearly, further studies are also needed to identify the most suitable gesture sets for different tasks.

## 6 Conclusion

In this paper, we proposed a novel approach to interact with video environments in an immersive and intuitive way. Using their avatar, users can move inside the footage in three dimensions, and place virtual objects inside the scene depicted on the video. Knowledge about the height of the user and the layer model enable the system to place the video avatar in three dimensions. The system can be used for various applications, for example, the prototyping and evaluation of ubiquitous and situated systems. We presented an example scenario, where designers first placed an interactive public display in a video environment and a stakeholder then explored the resulting scenario.

The initial prototype, though limited, used web technologies to illustrate the feasibility of the approach. A first study provided initial evidence for a high degree of immersion and the usability of the proposed approach. Future work will focus on carrying out a series of further user studies. With the latter, we plan to explore properties of mirror image avatars as a means of interaction in simulated environments, to compare this approach to alternatives, e.g., 3D controllers, and to investigate the integration of mobile devices with the system, e.g., as a secondary controller for content selection.

## References

1. Ahn, S.C., Lee, T.-S., Kim, I.-J., Kwon, Y.-M., Kim, H.-G.: Large display interaction using video avatar and hand gesture recognition. In: Campilho, A.C., Kamel, M.S. (eds.) ICIAR 2004. LNCS, vol. 3211, pp. 261–268. Springer, Heidelberg (2004)
2. Ballendat, T., Marquardt, N., Greenberg, S.: Proxemic interaction: designing for a proximity and orientation-aware environment. In: Proceedings of ITS 2010, pp. 121–130. ACM (2010)
3. Benko, H., Wilson, A.D.: Multi-point interactions with immersive omnidirectional visualizations in a dome. In: Proceedings of ITS 2010, pp. 19–28. ACM (2010)
4. Gotardo, P.F.U., Price, A.: Integrated space: authoring in an immersive environment with 3D body tracking. In: SIGGRAPH 2010 Posters. ACM (2010)
5. Iwabuchi, E., Nakagawa, M., Siio, I.: Smart makeup mirror: computer-augmented mirror to aid makeup application. In: Jacko, J.A. (ed.) HCI International 2009, Part IV. LNCS, vol. 5613, pp. 495–503. Springer, Heidelberg (2009)
6. Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A., Fitzgibbon, A.: Kinectfusion: real-time 3D reconstruction and interaction using a moving depth camera. In: Proceedings of UIST 2011, pp. 559–568. ACM (2011)
7. Morikawa, O., Maesako, T.: Hypermirror: toward pleasant-to-use video mediated communication system. In: Proceedings of CSCW 1998, pp. 149–158. ACM (1998)
8. Nancel, M., Wagner, J., Pietriga, E., Chapuis, O., Mackay, W.: Mid-air pan-and-zoom on wall-sized displays. In: Proceedings of CHI 2011, pp. 177–186. ACM (2011)
9. Särkelä, H., Takatalo, J., May, P., Laakso, M., Nyman, G.: The movement patterns and the experiential components of virtual environments. *Int. J. Hum.-Comput. Stud.* **67**(9), 787–799 (2009)
10. Andrés del Valle, A.C., Opalach, A.: The persuasive mirror: computerized persuasion for healthy living. In: Proceedings of HCI International 2005 (2005)
11. Vera, L., Gimeno, J., Coma, I., Fernández, M.: Augmented mirror: interactive augmented reality system based on kinect. In: Campos, P., Graham, N., Jorge, J., Nunes, N., Palanque, P., Winckler, M. (eds.) INTERACT 2011, Part IV. LNCS, vol. 6949, pp. 483–486. Springer, Heidelberg (2011)
12. Vogel, D., Balakrishnan, R.: Distant freehand pointing and clicking on very large, high resolution displays. In: Proceedings of UIST 2005, pp. 33–42. ACM (2005)

# Sci-Fi Gestures Catalog

## Understanding the Future of Gestural Interaction

Lucas S. Figueiredo<sup>(✉)</sup>, Mariana Pinheiro, Edvar Vilar Neto,  
Thiago Chaves, and Veronica Teichrieb

Federal University of Pernambuco, Recife, PE 50740-560, Brazil

{lslf, mgmp, excvn, tmc2, vt}@cin.ufpe.br

<http://www.cin.ufpe.br/voxarlabs>

**Abstract.** In Science Fiction (Sci-Fi) movies, filmmakers try to anticipate trends and new forms of interaction. Metaphors are created allowing their characters to interact with futuristic devices and environments. These devices and metaphors should be target of research considering they have proven to be useful before. Moreover, the impact of the new interfaces on the audience may indicate their expectations regarding future gesture interactions. Thus, the first goal of this work is to collect and expose a compilation of gestural interactions in Sci-Fi movies, providing a catalog to researchers as resource to future discussions. The second goal is to classify the collected data according to a series of criteria. The catalog is also open to new content contribution, and fellow researchers are invited to provide additional entries of hand gesture scenes from any Sci-Fi title as well as suggestions about new classification criteria and amendments on the already provided content.

**Keywords:** Sci-Fi movies · Hand gestures · Gesture interaction · User experience

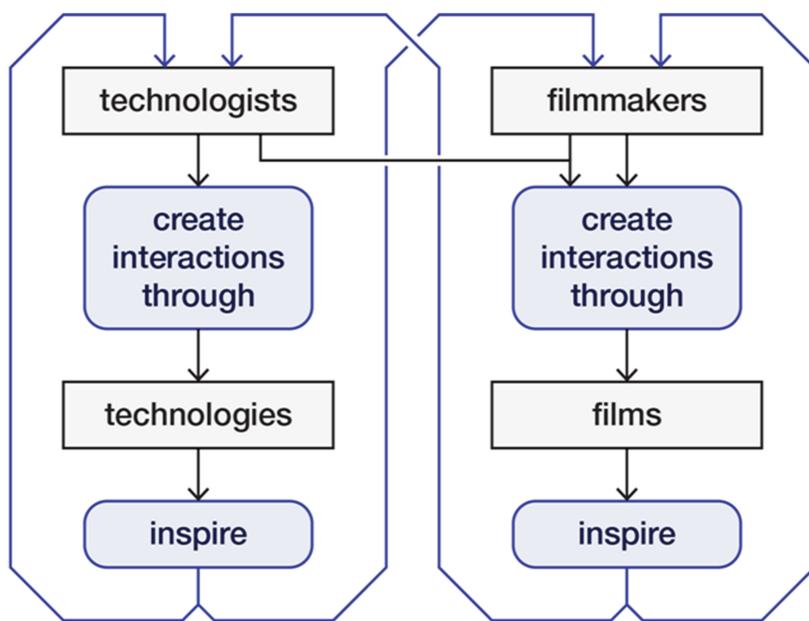
## 1 Introduction

Movies are a kind of entertainment that has a high impact in the formation of the general public mindset. In particular, Science Fiction (Sci-Fi) movies try to understand and anticipate trends of the future mainly related to new technologies [1]. Some producers and directors often use emerging interaction paradigms, such as gesture-based interaction, to create a plausible vision of the future.

Given that moviemakers have resources and freedom to create characters, stories and scenarios, without common limitations of the real world, Sci-Fi movies can present a particular vision of the future that we are not familiarized with. This can help to make new technologies and interaction models widely known to the general public, contributing to popularize their adoption and to highlight aspects that can be improved by the industry and academy.

On the other hand the Sci-Fi media has potential to reveal or emphasize research trends regarding particular interaction paradigms, devices and interfaces for specific tasks and application domains. Once particular visions of filmmakers are well accepted by the audience, the research on the same topic is boosted by additional motivation, the

upcoming technology starts to be part of public's imaginary and an inner curiosity grows with questions like "how would this work in real life?". Figure 1 shows the influence cycle between the filmmaking industry and the interaction designing community. At first, there is the local influence cycle, which makes explicit that technologists and filmmakers end up influencing themselves. The cross influence cycle is also possible, when one side influences or inspires the other through their products (technologies or films). In addition, as suggested in [1], the flow of influence between technologists and filmmakers can potentially create a collaboration environment in which both sides can create together new interactions and present them in movies.



**Fig. 1.** Filmmaking industry and interaction designing community influence flow

In cases that producers and directors are aided by interaction designers the movie can be seen as a powerful tool for the designer to explore a new concept, envisioning with high visual fidelity how the interface should work. In this case it should be taken into account that there is a certain amount of influence of the entertainment industry over the designer vision. A set of additional concerns like the visual appeal of the scene, the plot development and how the character will look like while using the new interface which may not intersect with goals like realizing the targeted task efficiently or providing a good experience for the user.

Thus, the movies vision of Future User Interfaces (FUIs) often says more about us and the characters than they do about the future [2]. In others words, Sci-Fi movies end up adapting technologies to the characters characteristics. This point of view can be explored on future user analysis, as for example examining reactions of potential users

while watching FUIs being used by the characters and understanding their expectations regarding the same FUIs in the real world. Design problems can be anticipated by gathering feedback from those potential users about the applicability of a particular FUI, understanding if they fit in the specific application domain or, for example, if the input device (e.g. gloves) is accepted on the aimed context.

Despite the industry influence on the FUI design, the scene content is valuable for a range of research purposes. Even if we do not know how the audience responds to a specific FUI, the design process can take it as a source of inspiration. This gives pre-made concepts to designers and developers to create solutions that can be both made with current technologies and be a start point to explore new concepts. Prototyping processes like the Wizard-of-Oz design can benefit from visuals to aid the setup definition including the particular task of designing FUIs which demands knowledge about similar interface concepts that have been designed before. With speculative design in mind, visualizing new concepts of interaction can emerge provocative questions, dialoguing with different ideas of future.

On the other hand, the state of art of hand tracking algorithms and devices is showing promising results [3, 4]. The industry is also presenting accessible devices which provide real time tracking results such as the Microsoft Kinect [5], Leap Motion [6] and Myo [7]. These devices represent a turning point on the design of gestural interfaces due to their low price and the minimum required hardware and setup of the scene broadening the range of gesture interfaces applications and demanding research on the gestural interaction design.

Considering these topics, this paper focuses on hand gestures interaction compiled from Sci-Fi movies. The main goal of this work is to provide an open catalog of these interactions on Sci-Fi scenes, empowering researchers with a tool to gather inspiration for the task of designing new interfaces as well as to perform analysis over the target content. In order to accomplish this we collected and categorized scene parts of hand gestures being performed on different titles. Moreover, our work examines specific aspects related to the interactions and, as for example, the used input and output devices, the performed gesture and the result task executed by the system, among others. Each gesture is tagged considering each one of the chosen criteria. This allowed us to verify supposed lessons from the cinema to designers and researchers as well as identify opportunities to maturing this area. Having this in mind, a web application is presented in order to make public the data from this work for academics as well as the general public. The application is a collaborative system, allowing visitors to contribute by increasing the data.

Supported by the current emerging devices and algorithms for hand tracking as mentioned before, in the first version of the catalog the scope is narrowed to arm and hand gestures. Moreover, currently the input data is gathered exclusively from movies and TV series in order to build a first set of video snippets. However the application may support content from other media such as video game scenes as well. Since the catalog targets the visual analysis of FUIs, it is required a small video for each entry, and so other art forms, such as comic books and novels, are out of scope.

As technical contributions there is the concept and implementation of the catalog as a tool for video analysis. The web application code is available and may be reused for similar video analysis of different data; the application is based on a Google

Spreadsheet content being adaptable to any type of categories as long as there is a video for each row entry. Secondly, video snippets containing hand and arm interactions were collected and categorized along 215 scene parts from 24 different titles. Figure 2 shows some of the scenes in which the users interact with the systems using gestures. In Ender's Game (Fig. 2a) and Sleep Dealer (Fig. 2c) the protagonists use their hands to remotely control a spaceship and an aircraft, respectively. The woman from Black Mirror (Fig. 2b) is playing a virtual violin, while Tony Stark from Iron Man (Fig. 2d) controls a hologram that represents the city and then he can zoom in and out any place of interest, or delete things from the model with just a flick. At last, in order to support additional analysis, all the scene parts were categorized according to some established criteria as for example the application domain, if there was an identified interaction pattern in the scene or which was the used input device.



**Fig. 2.** Examples of movies that compose the used dataset in this study

The main scientific contributions are related to the performed analysis and the used methodology. We found that the major part of the scenes do not fit on previous classifications proposed by the related literature. In some cases it is even difficult to relate the performed gesture to an accomplished task. This opens space for further investigations in order to understand the industry demands for interactions which are not yet established. Moreover, as an incremental contribution, we validate and extend the number of interaction patterns identified on the analyzed FUIs. As methodological contribution we present an end-to-end methodology to gather and display the material in order to perform video analysis on a target subject, including tools and resources. The result enables researchers to directly relate the video material with its classification as well as contribute to increase the gathered data. At last, we considered particular scenes containing the use of Telekinesis by superheroes which are not directly used for HCI purposes. However, in several cases, these gestures and their corresponding actions do fit in previous established patterns signalizing that there is a relation, mainly regarding manipulation tasks on gestural interfaces.

This paper is organized as follows: the next section presents works that also make use of information from Sci-Fi movies to perform analysis on various fields of knowledge. Section 3 discusses the methodology of the work. Section 4 presents the web application developed and Sect. 5 the results obtained. Section 6 presents a discussion about the results. Section 7 concludes the study and Sect. 8 makes a survey of possible future work that the outcome of the study can develop.

## 2 Related Work

Although this work is focused on gesture interactions in Sci-Fi movies, it represents only a subset in the space of human-computer interactions found in these films. Looking at different interactions from Sci-Fi movies, researchers have been dedicated to understand this new set of communications between human and machine. In one of his publications in 2012, “The Past 100 Years of the Future” [8], Aaron Marcus selected scenes from around 20 Sci-Fi movies along the last 100 years that contain communication between human-machine and described them according to the used interaction. His study is organized chronologically around the dates at which these movies were launched to the public. In some cases, he comments on the budget that the moviemakers had available to produce the scenes.

Schmitz et al. show a similar analysis [1], surveying ways of HCI from Sci-Fi movies, categorizing them according to their application domains and relating them to current technologies and prototypes under research. The authors indicate that movies can be a two-way road to HCI, i.e., they can anticipate trends and inspire future concepts of interaction and also collaborate with researchers and visionaries to the conception of scenarios using emerging concepts. The work suggests an influence flow between moviemakers and scientists regarding the use of HCI in movies, where producers and researchers can develop new ideas about future interactions (Fig. 1). Thus, technology can inspire movies in the same way that movies interpretation can give feedback about new concepts of devices and interactions. The authors discuss the inspiration that Sci-Fi movies pose to future technologies.

Shedroff and Noesel [9] contribute analyzing around 30 Sci-Fi movies and presenting lessons and insights about the interactions shown on them to HCI. As mentioned, Sci-Fi movies can take advantage to create their own vision because they do not need to limit themselves according to the current technologies. Particularly regarding the gesture-based interactions, the authors point to the existence of seven patterns that can be considered established in HCI. The identified patterns are represented as physical analogies from the real world, i.e., possess a sense of direct manipulation. In addition, this work shows that more complex gestures (that need to get support from GUI) can be difficult to be memorized and advice the use of others input channels to work around possible abstractions, such as speech interactions.

In [10], Christina York focuses on good practices for observation techniques aiming the creation of better interactions. For this purpose her work makes use of Sci-Fi media (e.g., scenes from the StarTrek movie). The analysis includes gestural interactions, as well as touch and voice interfaces.

Asian contributions in Sci-Fi movies are also reviewed by Aaron Marcus [11, 12]. According to the author, studying Chinese, Indian and Japanese productions shows more creations based on different cultures than properly a copy of western approaches. The survey collected in this work initiates a look at the differences and similarities among different cultures, contributing to extend the perspectives about the future of user experience design. Moreover it suggests that the metaphors, mental models, interactions and visual appeal on some titles can reveal cross-cultural influences.

Beside papers that study HCI in Sci-Fi films on a generic way, there are other studies on specific areas of knowledge seeking the correlation between the illustrated interactions and real life. For example, the work of Lorencik et al. [13] investigates the mutual influence of science fiction films in the field of Artificial Intelligence and Robotics. The scenes of six Sci-Fi movies were analyzed in relation to the systems that controlled the robots and the actions that the robots could perform. The work concludes that one of the main contributions of Sci-Fi movies to the fields of Robotics and AI is that they provide sources of inspiration about what can be done and also increase people interest in the area. The analysis performed by the authors focuses on the distinction of whether or not there is a real technology or a study on it.

Another example is the work of Boutillette et al. [14] that addressed the influence of Sci-Fi films in the development of biomedical instrumentation. In the work, more than 50 Sci-Fi movies were analyzed and, for each one, the scenes of instrumentation were extracted and analyzed in terms of knowledge and existing technology at that time. As a result, the films were divided into four categories in chronological order, and the analysis was made related to how the instruments were used, presented and whether it could be used at the time of the survey or in the future. At the end, a nonlinear relationship was found between the development of instrumentation and the ones shown in the Sci-Fi movies.

Although there are many studies regarding the relationship between movies and HCI, it is not possible yet to say which aspects can be incorporated or not in the design process. Even if producers and interaction researchers are working together, it is hard to understand which are the lessons to learn about this partnership. There are still opportunities to study the features of gesture-based interaction in this knowledge field.

### 3 Methodology

The aim of this work is to understand how the motion picture industry has been dealing with the hand gesture interaction paradigm as an insight of what can be built by tackling the previous problem of collecting and tagging the target data. The process can be divided into selecting the movies, searching for the gestural interactions within each of them, and then classifying it. Each of these steps is better explained in the following subsections.

#### 3.1 Data Collection

**Movies Selection.** The catalog concept is that it can grow as time passes by being open to anyone who wants to contribute adding new scenes, categories and more. This

way, the first set of selected movies aims to represent an initial set that is large enough to provide an analysis basis as well as a test set for the proposed tool. Since it is impossible for a small group to analyze all existing Sci-Fi movies, 25 volunteers were recruited by email. These volunteers were colleagues and they were asked to inform any movie they remembered that had any kind of gestural interaction using hands and/or arms. Together with the authors, there were a total of ten collaborators who indicated over 200 movies. These indications were checked to be Sci-Fi at IMDb site, which is a popularly known movie database. It was also checked whether they really had the kind of interaction this research was looking for. The remaining titles were restricted by their release date, i.e., the movies from 1995 until the date of this research were prioritized in order to reduce the scope of the work. This choice was made because newer movies are often inspired by the new emerging technologies, such as gesture-based interaction, giving them new insights to go even further. Additionally, there are already studies regarding the older ones (in the future, they will also be added in order to have a more complete database). Finally, the resulting subset contained a total of 24 movies.

**Gesture Search Procedure.** After the set of movies were defined, two of the authors were assigned to watch all the selected movies, each being responsible for half set, in order to identify the exact time frame the gestures occur during each movie. In order to analyze them all in a plausible time, they were watched up to 8x of the normal speed. For each scene with hand gesture found, the starting and ending time were annotated and filled in the Google Spreadsheet. With this in hand, a script was written (running the FFmpeg library [15]) in order to capture the scenes noted before. This allowed the scenes to be analyzed separately.

**Screening Criteria.** As mentioned before, each scene of interest was cut out of the movie for further analysis. Although all kind of hand interaction was collected, i.e., not only human-computer interactions, this allowed us to collect more kinds of metaphors used in gestures. The intention is to preferably allow false-positives rather than favoring the occurrence of false-negatives while segmenting the movies. Later on the scene parts were reviewed and the false-positives were filtered. Moreover, some non HCI interactions were selected, e.g., fighting scenes and Telekinesis interactions in super-heroes movies, as these scenes may serve as inspiration and be useful while considering the audience reaction to it.

**Analysis Criteria.** For this analysis, some aspects related to human-computer interaction were chosen. The first aspect to be analyzed was the relation between the action and executed task, i.e., the gesture performed in the scene and the consequent action realized by the system. In a first moment, the person who is analyzing only looks for the type of feedback passed to the user; the types of input and output devices used; and which kind of user the gesture was applied for. During the step of collecting parts containing gestural interaction, other data were becoming common between scenes, so it was decided to increase the amount of items for analysis. After a brief research about their validity, Telekinesis and Established Gestures (Pattern) were added as new items. The remaining categories are:

- Pattern: denotes if the gesture performed by the user and the corresponding action fit in a previous identified pattern (for example, “push to move” or “swipe to dismiss”).
- Feedback: denotes if the system provides any type of feedback and which type it is (e.g., “visual”, “audible” or “haptic”).
- Input Device: if the system presents clearly any required input device in order to track user actions (for example, “haptic gloves”).
- Output Device: denotes the device used for feedback (examples: “hologram”, “HMD”, “CAVE”, etc.).
- Domain: denotes the main application domain (examples include “medical”, “military” and “entertainment”).

The classification of each spotted gesture was done alongside cropping the videos. The researcher responsible for each scene filled a spreadsheet with the characteristics of the interaction according to the criteria discussed above. At completion, they reviewed together all the spreadsheet and discussed about the best classification whenever there was something they both didn’t agree with.

## 4 Web Interface

The gathered data was stored in its completion in Google Spreadsheets. By using tabletop.js javascript library it is possible to access the spreadsheet data and use it as database resource to the interface. It contains the selectable filters and the selected video snippets on its top part and the categories table at the bottom showing only the rows relative to the selected entries. The web application is available at <http://goo.gl/XSX5fn> and its source code is stored on a GitHub repository and can be found at <http://goo.gl/IpcAfl>.

We understand that to create an expressive compilation of Sci-Fi scene containing gesture interactions is an ambitious goal and our initial data-set represents a small sample of the target goal. Taking it into consideration we introduced the “Add a Scene” button on the top screen of the interface (Fig. 3), through which collaborators can send new entries of movies scenes containing human-computer gestural interactions. Each new entry is revised by the authors to check, among other details, if they fit in the Sci-Fi field, and then ported to the online data set.

Moreover, the web application can be useful for different data sets since it is almost entirely based on the data set stored in a Google Spreadsheet. By changing the source data set (by altering the targeted spreadsheet link) the web application adapts itself showing the new content in a similar way. The new columns will be categorized as filters and the videos from each row entry will be gathered and presented in the interface. This way, the interface is replicable and can serve to other focuses as long as they relate to video analysis in some way.

**SCI-FI GESTURES CATALOG**
[Add a Scene](#)

<b>Pattern</b> <input type="checkbox"/> Extend the hand to shoot <input type="checkbox"/> Knock-knock to turn on <input type="checkbox"/> Pinch and spread to scale <input type="checkbox"/> Point or touch to select <input type="checkbox"/> Push to move	<b>Type</b> <input type="checkbox"/> Detic <input checked="" type="checkbox"/> Emblems <input checked="" type="checkbox"/> Iconic <input checked="" type="checkbox"/> Metaphoric	<b>Manipulation</b> <input type="checkbox"/> No <input checked="" type="checkbox"/> Yes	<b>Telekinesis</b> <input type="checkbox"/> No <input checked="" type="checkbox"/> Yes	<b>Year</b> <input type="checkbox"/> 70 <input type="checkbox"/> 1889 <input type="checkbox"/> 1940 <input type="checkbox"/> 1963 <input type="checkbox"/> 2000 <input type="checkbox"/> 2008 <input type="checkbox"/> 2009 <input type="checkbox"/> 2010 <input type="checkbox"/> 2011	<b>Feedback</b> <input type="checkbox"/> Audible <input type="checkbox"/> Haptic <input checked="" type="checkbox"/> Visual	<b>Input-device</b> <input type="checkbox"/> Attached cables <input type="checkbox"/> Gloves with markers <input type="checkbox"/> Hand tracking sensor <input type="checkbox"/> Haptic gloves <input type="checkbox"/> Mug <input type="checkbox"/> Pen <input type="checkbox"/> Skin properties	<b>Output-device</b> <input type="checkbox"/> Analog pointer <input type="checkbox"/> Cave <input type="checkbox"/> Contact lens <input type="checkbox"/> Digital table <input type="checkbox"/> Hmd <input type="checkbox"/> Hologram <input type="checkbox"/> Monitor <input type="checkbox"/> Projection <input type="checkbox"/> Robot	<b>Domain</b> <input type="checkbox"/> Construction <input type="checkbox"/> Cooperative <input type="checkbox"/> Domotic <input type="checkbox"/> Entertainment <input type="checkbox"/> Information technology <input type="checkbox"/> Medical <input type="checkbox"/> Medicina <input type="checkbox"/> Military
--	--	---	--	--	--	--	---	---

[Refresh](#)







[<>](#) [9](#) [10](#) [11](#) [12](#) [13](#) [>](#)

Index	Title	Start	End	Pattern	Type	Manipulation	Telekinesis	Year	Feedback	Input-device	Output-device	Domain
1	X-Men 2000 1080p BrRip x264 YIFY.mp12	1:13:54	1:14:02	-	Emblems	No	No	2000	Haptic	-	-	-
2	X-Men 2000 1080p BrRip x264 YIFY.mp16	1:25:41	1:25:45	-	Emblems	No	Yes	2000	Haptic	-	-	-
3	Johnny Mnemonic (1995) [Extended Cut].avi	1:07:39	1:08:07	-	Emblems	No	No	2021	Visual, Audible	Hand Tracking Sensor	Monitor	Cooperative

**Fig. 3.** Web application implemented interface

## 5 Results

The following subsections will receive the name and title of each category concerned and will explain the results obtained for each one of them.

### 5.1 Movies Segmentation and Classification

After selecting 24 Sci-Fi movies, a total of 219 different excerpts containing hand gestures were extracted from them. Each segmented scene was classified in categories

according to aspects related to human-computer interaction that will be addressed in the next subsections.

In a first moment, the person who is analyzing only looks for the type of feedback passed to the user; which kind of user the gesture was applied for; and the types of input and output devices used. During the step of collecting parts containing gestural interaction, other data were becoming common between scenes, so it was decided to increase the amount of items for analysis. After a brief research about their validity, Telekinesis and Established Gestures were added as new items.

**Patterns.** Some pattern classifications regarding gestures are already defined in literature, for example the Shedroff et al. [9]. They conducted a major study of Sci-Fi movies in the HCI field regarding many aspects. The classification highlights a basic gesture vocabulary commonly used in these types of movies. For this category (Patterns) a benchmarking using this work was conducted and the classification of gestures established by them was applied because they standardize hand gestures used in Sci-Fi movies, which fits in this work. The classification divides the gestures into seven types relating action and result:

- Wave to activate
- Push to move
- Turn to rotate
- Swipe to dismiss
- Point or touch to select
- Extend the hand to shoot
- Pinch and spread to scale.

Another pattern classification, is suggested by the work of Wobbrock et al. [16]. But this work is related only to touch surfaces, therefore, initially, it was decided not to include their definitions into the classification. However, during the analysis of the scenes, it was noticed that one of Wobbrock's pattern definition fits with a gesture found in two scenes from two different movies, thus, it was further included in this work. This pattern, named "Knock-Knock to Turn On", denotes gestures with any part of the user's hand or arm touching twice the target system surface.

During the categorization process 37.83 % of the analyzed scenes contained some gesture that could be embedded into this classification, namely the vast majority of gestures found were too complex or not established in the classification of Shedroff et al. Regarding the scenes that fit in one of the categories there were a predominance in the use of gestures "Push To Move" and "Swipe to Dismiss", occurring on 11.41 % and 10.04 % of the scenes, respectively, as can be seen in the Table 1.

Another observation found is that some gestures, categorized or not, were used for completely different tasks in the same or different movies. When they occurred in the same movie, it was often a superhero film, mainly because the filmmakers are able to use the freedom that superheroes provide of moving things using the power of mind, or Telekinesis (to be discussed in Sect. 6.3). There is a part of the uncategorized gestures related to specific activities that humans are used to perform daily and these gestures were, mostly, exactly as it would be done in the real world turning difficult the task of create a really perfect pattern for them, since all human actions have at least one

performance standard that all the people do in the same way. The other part is composed of complex gestures using intense arm and hand movements without a particular pattern. Because of this, is considered a good result that more than 30 % of the gestures found match with one of the suggested patterns.

**Table 1.** Frequency of each one of the interaction patterns found in the scenes of this survey.

Identified interaction pattern	Percentage of occurrences
Push to move	11.41 %
Swipe to dismiss	10.04 %
Point or touch to select	5.47 %
Pinch and spread to scale	4.56 %
Turn to rotate	4.10 %
Knock-knock to turn on	0.9 %
Wave to activate	0.9 %
Extend the hand to shoot	0.45 %
None	62.17 %

**Input and Output Devices.** By considering the input and output devices, this work allows a notion of future technologies. The fact that one particular device appears in movies facilitates the audience acceptance of similar devices and understanding of the interactions, reducing the impact of its reception if at some point it becomes a real product. Table 2 shows the values of the occurrences for input devices.

About the input devices found, it can be seen in Table 2 that the most common case was the absence of identification of any kind of input device (70.81 % of the scenes). It reveals the idea that the interaction will be so natural that we will not need any physical device to interact with. In the cases a device was needed, the most used were gloves, optical wireless and haptic (21 %), and it shows that filmmakers suppose natural interactions should be used even if a device is needed. Summarizing, both approaches try to predict that future interfaces will be as natural as possible, not needing any device or complex interaction, using only natural gestures that users are used to perform.

The output devices have a wider variety, as shown in Table 3, but coincide with what, today, are considered high-tech display environments, such as HMDs and CAVE environments. There are still devices that are not so high-tech, as for example TVs. As can be seen, the output device often proposed by filmmakers, but not yet implemented, is the volumetric projection or, as it is commonly known, the hologram projection. Found in 31.96 % of the scenes, this device is a strong proof of the influence of films in the research in HCI as several studies on the subject can be found in the literature, for example, [17–19]. It is also a supposition of what the filmmakers guess the society expects from new output devices, a new and exciting way to experience virtual contents. The hologram has been studied as a substitute for physical prototypes and as an alternative to 3D displays. However, when the topic is interaction, researchers have different opinions and, until now, the main works converge that it is not possible to interact with holograms directly. In resume, the hologram has been generally used to

insinuate a high-tech vision of the future and the interactions have been performed in a straightforward manner, as if a physical object was being manipulated by the user.

**Table 2.** Proportion of different input devices found in the scenes.

Input devices	Percentage of occurrences
Optical wireless gloves	17.35 %
Cables plugged into the body	5.47 %
Haptic glove	3.65 %
Motion sensors	0.91 %
Pen	0.91 %
Biological recognition control	0.45 %
Mug	0.45 %
None	70.81 %

**Table 3.** Proportion of different output devices found in the scenes.

Output devices	Percentage of occurrences
Hologram	31.96 %
Transparent screen	19.63 %
Monitor	13.24 %
CAVE	4.56 %
HMD	4.56 %
Contact lens	4.1 %
Digital table	0.91 %
Projection	0.91 %
Analog pointer	0.45 %
Robot	0.45 %
Shapeshifting display	0.45 %
Water Tap	0.45 %
None	18.33 %

**Application Domain.** Within the analyzed data set, the predominant application domains were military, public security, corporative and domotic (Table 4). In addition to these domains, it was also perceived the targeting of HCI applications in the areas of IT, entertainment and medical applications. This guidance allows the estimative of the main areas that filmmakers are directing their attention regarding FUIs. Besides, it, points to areas that are receiving more attention from the industry innovation departments. The appearance of FUIs in movies acts like a preliminary test of acceptance by potential users. Having people getting used to see the FUIs in the movies makes it easier to begin their insertion in some application domains.

The application domains found in the movies of the database are of high importance for society. As it can be seen in Table 4, the first 4 most covered domains may affect the majority of people. In the public security area, for example, our second most

covered domain, there is a growing number of studies that point to the use of the human body for safety increase, e.g., the use of fingertips [20] or eye-tracking [21] to access or activate some systems. It is known that in the home automation area there are many researches about how to make it more effective and easy to use robots or machines so that they only need a few commands to perform a task. Some examples can be seen in [22] who have made an extensive study of the domotic area using gesture-based interaction since the state of art of gestural interfaces for intelligent domotic until socio-technical aspects of gestural interaction. These interactions were commonly seen in Sci-Fi movies, including the ones produced in 1995, what shows us that movies have been influencing research fields with their visionary interfaces and interactions.

**Table 4.** Percentage of gestural interactions for each one of the found application domains.

Application domain	Percentage of occurrences
Public security	17.35 %
Military	16.43 %
Corporative	10.95 %
Domotic	10.5 %
Information technology	10.04 %
Entertainment	9.58 %
Research	8.21 %
Construction	1.37 %
Medical	1.37 %
Robotic	0.9 %
Aerospace	0.45 %
Musical	0.45 %
Not defined	12.4 %

## 6 Discussion

After creating a movie scenes collection composed of more than two hundred scenes from 24 Sci-Fi movies, these scenes were categorized according to five criteria: established gestures, feedback, input and output type of device and application domain. In the following subsections we discussed about the results found in each of these criteria.

### 6.1 Focused Analysis

As shown in Table 1, the set of gestures considered established were confirmed in our analysis. In total, 81 gestures were categorized into this set, being 11.41 % of them classified as “Push to Move” followed by “Swipe to Dismiss” (10.04 %). All the others gestures were found in our review. “Wave to Activate”, “Knock-Knock to Turn On” and “Extend the Hand to Shoot” were the least found. On the other hand, we found at least an indication of another pattern in gesture-based interaction. The interaction here

named “Knock-Knock to Turn On” is present in two scenes from two movies. This action/task relation may already be found in the real world, for instance in the LG G2 phone, in which the user can touch the screen twice to turn the screen on [23]. This gesture is also described as a tabletop interaction in the Wobbrock’s work [16], which reinforces that it can be considered an established interaction pattern. In spite of that, the majority of gestures gathered in the present paper seem to be way more complex and abstract than those classified into the pattern category. It is not possible to say that there are no precedent occurrences of these gestures either in the academy or in the industry, though. That said, a more in-depth analysis regarding gestures classification must be done to find out whether the filmmaker’s vision is plausible or aligned with current researches at least. Moreover, even if these gestures are unprecedented in the literature, a public survey can be done to get feedbacks from the general public in order to figure out their applicability and their potential to be studied and carried out.

Looking at the environment and purpose of the gestures – the Application Domain – most of them were based on Military and Public Security applications. Corporative environment, Domotic (including home automation and similar), Information Technology, Entertainment and Scientific Research were often shown in this review. Medical, Construction and Robotics come next. Musical applications were also listed. The frequency in which each application domain appears is related to the movie’s theme, but the fact they appeared possibly means they have a high potential to be explored within this interaction paradigm.

Regarding the types of input shown in the listed gestures, although Optical Marker Gloves, Body Plugs and Haptic Gloves appear in many scenes, they belong to, respectively, Minority Report, Sleep Dealer and Johnny Mnemonic. A lot of other devices were found, including Motion Sensor, Pen, Staff and a Mug. This variety of devices may occur due to the Sci-Fi nature of presenting their own vision of the future. However, most of the interactions were performed without any kind of support device to promote input. In this case, this absence of input devices may be a trend to seek more natural interaction interfaces. A special attention should be given regarding this type of interaction, since there are a lot of Sci-Fi movies that involve ETs and their hypothetical high advanced technologies, there is also a new whole group of gestures and interactions they use to control spaceships, weapons, among others, that should be observed. For example, Marcus A. [8] already discusses this topic and compares the way a three fingered alien, acted by Sharlto Copley in District 9, controls a spaceship to the way Tom Cruise, in Minority Report, interacts to the Precog scrubber interface. The author affirms that in District 9 the alien creature controls the spaceship interacting with its interface in an elegant and fast-paced way that is much more beautiful and fluid than the interaction done by Tom Cruise’s character.

Hologram has been explored a lot as an output device being found in 31.96 % of the scenes and 11/24 of the movies. It is followed by Transparent Screen (19.63 % of the scenes) and Monitor (13.24 %). Other output devices include Water Tap, HMD, Shapeshifting Display, Robot, Projection, Digital Table, Contact Lens, Analog Pointer and CAVE. Both Hologram and Transparent Screen are technologies currently under research, i.e., we do not find them in the market yet. One reason they appear often in this review is because moviemakers are connected to researchers and developers in order to create a plausible scenario of the near future.

## 6.2 Use of Telekinesis and Superhero Cases

In addition to Sci-Fi, we also watched two superhero and one alien movies. These movies have scenes with a type of Telekinesis interaction supported by hand-based gestures. We defined Telekinesis as “the power to move objects from a distance without physical contact”. That means some characters have the ability to move matter using their hands as a kind of pointer. We collected these scenes because we believe these interactions may show us gestures beyond the usual paradigm. If we are able to understand how to make this type of interaction feasible, we can appropriate the movie’s vision to start new ideas and researches.

In many cases the hand gestures from these movies have no relation to computer interfaces. In these cases they do not have any input or output devices, since the most gestures are performed in fight scenes and use Telekinesis approach. Nevertheless, some gestures use established pattern of gesture-based interface, for instance, “Extend the Hand to Shoot” and “Push to Move”. However, we have the opportunity in these movies to look forward to new forms of interaction that we are not familiarized with. Thus, these scenes may provide us a rich source of inspiration to create innovative solutions.

## 6.3 Methodological Challenges

The large amount of Sci-Fi movies demands a lot of time to extract the gestural interaction scenes. In this survey 24 movies were analyzed, that means watching them, extracting the exact timestamp of the desired scenes and fit the interaction into the defined criteria to add the appropriate tags. So, a challenge is to continue expanding this database of scenes, i.e., to include movies that were published after this research and also those ones that were not included because of the restriction in the scope of the research.

## 7 Conclusion

This paper sought to collect hand gestures from Sci-Fi movies aiming to classify and catalog them in order to make a database to future works. From 24 movies, 219 scenes were captured and classified according to five criteria that take into account movies approaches and interfaces present in each interaction. Moreover, with this result in hand, a web application was developed in order to make all the results available to anyone. These results give pre-made concepts to designers and developers and they are a source of inspiration to create solutions that can be both made with current technologies and be a starting point to explore new concepts. Besides, the visualization of new concepts of interaction can emerge provocative questions, dialoguing with different ideas of future.

Furthermore, we were able to notice that some previous results from the literature were confirmed in our study. We also noticed the gesture-based paradigm is maturing since we could observe previous indications of other emerging patterns as the gesture patterns suggested by Shedroff and Noessel in [9] and the “knock-knock” pattern perceived during the development of this study.

Simultaneously, we have found some shortcomings in our analysis. Some criteria were flawed and we decided to take them out of this review. More work is needed to correct them. Analyze others data, such as gesture classification, was in our previous goals and it will be added in future studies. In addition, more research must be conducted, analyzing possible correlations between criteria and extending the previous results.

## 8 Future Work

As future works, further analysis about the data collected in this survey should be made. It is also intended to combine columns of the scene database, and observe trends and similarities with what is happening in the real world. For example, it could be done a cross-cultural study, i.e., to watch and observe patterns of similarity or differences between gestures across different cultures, namely western/eastern or to notice patterns from each continent.

Besides, the database should be incremented with more movies and more categories of analysis. In this initial work, there were only seven categories but it is already in research the addition of more. For example, some simpler categories detailing the year of launch of the title as well as its production region (in order to perform cross-cultural analysis). Other more elaborated criteria are also target for future work such as the feasibility of the gestures or whether the gestures already exist in the real world. Also, not only the current database should be expanded but it may also include other sources of material aside of movie scenes, for example, video game scenes, cartoons, among others.

At last, the portal will demand additional solid crowdsourcing mechanics to encourage collaborators to add, modify and remove content in a safe and coherent way. The use of pre-analysis tools for dynamic chart generations, and the display of selective table content are also in the scope for future improvements. Finally, a video compilation module is desired to create a single video containing all the filtered scene parts enabling a more direct analysis of the target content.

## References

1. Schmitz, M., Endres, C., Butz, A.: A survey of human-computer interaction design in science fiction movies. In: 2nd International Conference on INtelligent TEchnologies for Interactive enterTAINment, Cancun, vol. 7, pp. 1–10 (2007)
2. Leap Motion: how do fictional UIs influence todays motion controls? <http://bit.ly/1sV4NOC>
3. Oikonomidis, I., Kyriazis, N., Argyros, A.: A efficient model-based 3D tracking of hand articulations using kinect. In: BMVC, Dundee, vol. 1, p. 3 (2011)
4. Stenger, B., Mendonça, P.R., Cipolla, R.: Model-based 3D tracking of an articulated hand. In: Computer Vision and Pattern Recognition, vol. 2, pp. 310–315. IEEE Press, Kauai (2001)
5. Microsoft kinect for windows. <http://www.microsoft.com/en-us/kinectforwindows>

6. Leap Motion Mac and PC motion controller for games, design, and more. <https://www.leapmotion.com>
7. Thalmic Labs Inc.: Myo - gesture control armband. <https://www.thalmic.com/en/myo>
8. Marcus, A.: The past 100 years of the future: human-computer interaction in science-fiction movies and television. [http://www.amanda.com/wp-content/uploads/2012/10/AM+A.SciFI+HCI.eBook\\_LM10Oct12.pdf](http://www.amanda.com/wp-content/uploads/2012/10/AM+A.SciFI+HCI.eBook_LM10Oct12.pdf)
9. Shedroff, N., Noessel, C.: *Make It So: Interaction Design Lessons from Science Fiction*. Rosenfeld Media, New York (2012)
10. York, C.: To boldly go where no observer has gone before: sci-fi movies for UX practice. In: User Experience Magazine, vol. 13, p. 2 (2013)
11. Marcus, A.: User-experience and science-fiction in Chinese, Indian, and Japanese films. In: Marcus, A. (ed.) DUXU 2013, Part II. LNCS, vol. 8013, pp. 72–78. Springer, Heidelberg (2013)
12. Marcus, A.: On the edge: beyond UX in western science fiction. In: User Experience Magazine, vol. 11, p. 30 (2013)
13. Lorencik, D., Tarhanicova, M., Sincak, P.: Influence of sci-fi films on artificial intelligence and vice-versa. In: IEEE 11th International Symposium on Applied Machine Intelligence and Informatics (SAMI), Herl'any, pp. 27–31 (2013)
14. Boutillette, M., Coveney, C., Kun, S., Menides, L.: The influence of science fiction films on the development of biomedical instrumentation. In: Bioengineering Conference, Hartford, pp. 143–144 (1999)
15. Bellard, F., Niedermayer, M., et al.: Ffmpeg. <http://ffmpeg.org>
16. Wobbrock, J.O., Morris, M.R., Wilson, A.D.: User-defined gestures for surface computing. In: SIGCHI Conference on Human Factors in Computing Systems, pp. 1083–1092. ACM, Boston (2009)
17. Opiyo, E., Horváth, I., Rusák, Z.: Investigation of the scope of support provided by holographic displays in conceptual design. In: Talaba, D., Amditis, A. (eds.) Product Engineering, pp. 353–365. Springer, Dordrecht (2008)
18. Lee, H.: 3D holographic technology and its educational potential. In: TechTrends, vol. 57, pp. 34–39. Springer, New York (2013)
19. Bimber, O.: Combining optical holograms with interactive computer graphics. In: ACM SIGGRAPH 2005 Courses, p. 4. ACM, Los Angeles (2005)
20. Jing, P., Yepeng, G.: Human-computer interaction using pointing gesture based on an adaptive virtual touch screen. Int. J. Signal Process. Image Process. **6**(4), 81–92 (2013)
21. Cantoni, V., Galdi, C., Nappi, M., Porta, M., Riccio, D.: GANT: gaze analysis technique for human identification. Pattern Recogn. **48**, 1027–1038 (2014)
22. de Carvalho Correia, A.C., de Miranda, L.C., Hornung, H.: Gesture-based interaction in domotic environments: state of the art and HCI framework inspired by the diversity. In: Kotzé, P., Marsden, G., Lindgaard, G., Wesson, J., Winckler, M. (eds.) INTERACT 2013, Part II. LNCS, vol. 8118, pp. 300–317. Springer, Heidelberg (2013)
23. User Guide LG G2. LG. <http://goo.gl/Mei0d5>

# TV Interaction Beyond the Button Press

## Exploring the Implications of Gesture, Pressure and Breath as Interaction Mechanisms for a TV User Interface

Regina Bernhaupt<sup>1</sup>, Antoine Desnos<sup>1</sup>, Michael Pirker<sup>2(✉)</sup>, and Daniel Schwaiger<sup>2</sup>

<sup>1</sup> ICS team, IRIT, Université Paul Sabatier, Toulouse, France  
`{regina.bernhaupt, antoine.desnos}@irit.fr`

<sup>2</sup> User Experience Research, Ruwido Austria GmbH, Neumarkt, Austria  
`{michael.pirker, daniel.schwaiger}@ruwido.com`

**Abstract.** In order to enhance users' interactions with TV user interfaces we developed a prototypical multimodal interaction mechanism that combines tilting, pressing and puffing as input modalities for a novel interface. The interaction mechanism has been evaluated in an exploratory user experience and usability study that used a within subjects design investigating tilt as input mechanism to navigate through the 3D interface compared to tilt combined with pressure and breath input. Results of this first exploratory study indicate that while this uncommon and unfamiliar way to interact with a novel TV user interface impacts usability scores which were below average compared to traditional remote controls, the user interface approach in combination with the new interaction modalities resulted in above-average scores for the user experience dimension of hedonic quality. The findings are subsequently reflected and implications of using alternative input modalities for TV user interfaces are discussed.

**Keywords:** Remote control · Breath · 3D position · Gesture · TV · User interface · UI · User experience · Usability

## 1 Introduction

Today TV user interfaces present more content and functions than ever before. Current offers enable users to watch hundreds of TV channels, provide an electronic program guide (EPG), and include a variety of apps or functions allowing to buy video on demand (VoD) movies and series, to record programs or to time-shift a program. To help users to interact with this multitude of services, the TV user interface has to provide an easy and intuitive menu structure as well as interaction mechanisms that can deal with this multitude of content and functionality, while at the same time ensuring to provide a good user experience (UX). While limitations of standard remote controls have been reported in detail (e.g. [5]), interacting with a TV is still associated with the use of a ordinary infra-red remote control by the majority of people, although also the remote control saw a variety of changes over the past 50 years, including the extension of number of buttons,

possibilities for text entry, button reduction in conjunction with on-screen UIs, and the usage of modalities like touch or speech.

The general goal of our research is to investigate new forms and combinations of modalities to enhance TV user interfaces beyond standard grid-based structures and to find an interaction mechanism that is providing a novel and positive user experience. In this paper, we present results of a study that uses gesture as a possible means for playful interaction in the context of IPTV. Our research hypotheses were based on the idea that through the use of a playful approach with multimodal input (that is also used by e.g. gaming devices like the Nintendo Wii or Microsoft X-Box Kinect), a positive impact on the UX should be observable. Furthermore, we wanted to investigate the usability of gestures, especially accelerometer based gesture (tilt), as well as other modalities like pressure (press) and noise/voice input (puff) for TV interaction.

## 2 Related Work

### 2.1 Gesture Interaction and Interactive TV

Von Hardenberg and Bérard [3] discuss three types of requirements for a system using gesture interaction: the ability to detect or recognize gesture interaction, the identification (which type of object from a certain class is present – e.g. the recognition of certain movement and gestures), as well as the capacity to track gestures.

Gesture is widely accepted as a possible means of controlling devices due to its success in video game consoles like the Nintendo Wii and Microsoft Xbox Kinect. When gestures are applied in a TV environment, it offers the advantages that it does not require visual attention from the user on the remote control, but supports a completely blind usage, enabling the user to solely focus on the TV screen, and is reported to be well accepted by users [7, 9]. Gesture interaction based on a Wii-mote has been investigated in the context of TV [1], replacing traditional remote controls. Gestures seem appropriate to enable eyes-free interaction, and to avoid the continuous problem of selecting buttons on a standard remote control, although typing on a physical remote control will remain faster than using gesture to control the TV [1]. Vatavu [11] proposed a set of commands for interacting with augmented TV environments using gesture with a Wii controller. It was observed that the majority of participants preferred buttons to motion gestures, especially for abstract tasks (e.g. mute or menu), and that purely gestural techniques tend to be slightly lower in terms of performance. Gesture was also used in combination with mobile phones and real 3D TV simulations to investigate new forms of control on the mobile phone screen [4], differing from the work reported in this paper as two screens were involved.

### 2.2 Pressure, Deformation and Noise as Input

Hoggan et al. [6] dealt with the question whether squeezing is an effective input modality of mobile devices and if tactile feedback improves performance, as well as the effectiveness of squeezing as a manipulative interaction technique for use in mobile devices. Results for menu selection tasks show that squeezing is significantly faster than tilting, with and without tactile feedback, while both conditions facilitate successful interaction.

Not only voice, but also noise has been used to interact with user interfaces. Especially breath has been used by several researchers to implement interaction with interactive applications. Patel and Abowd [8] presented BLUI, a localized blowable user interface that allowed hands-free interaction via blowing at a laptop or computer screen to directly control certain interactive applications.

### 3 Problem Description

Our research goal and motivation was to investigate and get a deeper understanding of the usability and user experience when using accelerometer-based gesture (tilt) to navigate in a user interface that is not grid-based, but uses information representations in a pseudo 3D form of presentation. Based on recent positive reports from the general adoption of Wii-based interaction we wanted to investigate whether accelerometer-based gesture input results in the same Usability (measured with SUS questionnaire) and user experience (measured with AttrakDiff questionnaire) ratings than traditional TV interaction mechanisms (standard remote controls on standard IPTV UIs), as well as the implications on usability and user experience if the gesture input is enhanced with further input modalities (pressure, puffing). The comparison to traditional TV interaction mechanisms has been carried out by comparing the questionnaire scores to those from previous work in this field [10].

### 4 The Prototypical System

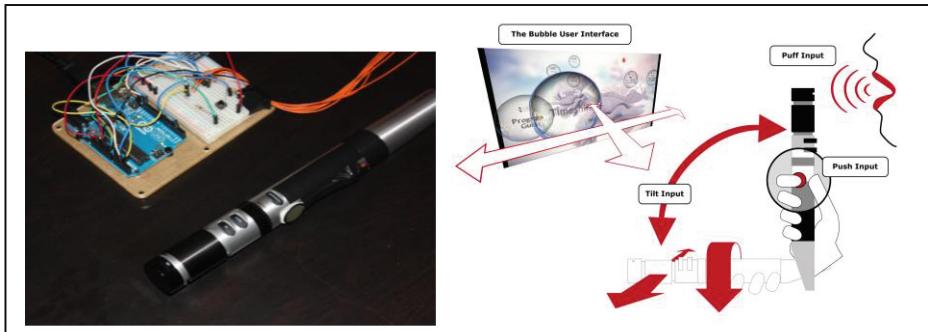
In the exploratory evaluation study, a prototypical user interface called Bubble UI and a remote control prototype described subsequently were used. The Bubble User Interface was designed with the goal to have an easy and intuitive way to navigate within large quantities of content, offering a novel and playful experience and improving the overall experience of the users. The UI represents the menus and content as a continuous and dynamic stream of information instead of static navigation and content elements. The representation of menus and content is carried out using dynamic round elements, which are visualized using the design metaphor of floating soap bubbles in the 3D space that can be directly manipulated in the synchronized user interface using the custom-tailored interaction mechanism. Further information on interaction, design rationale and further design aspects of the Bubble UI are discussed by the creators of the user interface in [2].

In order to complete our goal to offer a novel experience, we combined a set of analog inputs along with the GUI design. We chose Tilt, Press and Puff interaction as inputs because of both their analog nature and their close relation to the shapes and animations in the UI to ensure a consistent, novel and playful interaction ecosystem.

The remote control prototype which was used in the study was built using an Arduino Uno microcontroller board and sensors which were attached to an already existing remote control.

The gesture and the pressure modalities were implemented using the Arduino components, while the microphone used for voice was already included in the remote control that was used (cf. Fig. 1). The Arduino sensors we used in this case were a

standard force resisting sensor (FSR) and the ADXL335 accelerometer. The sensors were cased nicely in the remote control.



**Fig. 1.** Left image: The remote control prototype and the Arduino Uno Board; right image: Interaction with the Bubble UI: Possible movements were left/right and up/down.

We kept the Arduino board separated from the remote and hidden under a coffee table. The sensors were connected using one meter long cables. The accelerometer was glued inside the battery compartment to ensure that it will only move when the remote moves with the users hand. The FSR was attached to the side of the remote control and fixed in reach for left as well as right handed persons using insulating tape. The audio input from the built-in microphone controlled the puff input for accessing the main menu and the ‘back’ functionality. In the exploratory study, the navigation in the 3D space of the UI was carried out using the position-sensitive remote control prototype by tilting the remote control on two axes, forward, backward, left or right (see Fig. 1). In one of the study conditions – the enhanced condition - the prototype used a pressure-sensitive sensor that had to be pressed beyond a certain threshold for selecting an item (incl. visual feedback on force applied). The ‘puff’ functionality took over the metaphor of real soap bubbles and accessed the main menu (i.e. bubbles with menu items appeared on screen), or in the case of an already displayed menu hid the menu content (i.e. the user ‘blew’ the soap bubbles away).

## 5 Exploratory Evaluation Study

An experimental user study was conducted to investigate usability and user experience of the proposed user interface with accelerometer-based gesture input. The study followed a traditional usability study method including observation, interviews and domain-specific tasks to perform. Participants filled in the standard usability scale (SUS) questionnaire to measure usability, and the AttrakDiff questionnaire ([www.attrak-diff.de](http://www.attrak-diff.de)) to measure the perceived user experience with the dimensions of pragmatic quality, hedonic quality and attractiveness.

To limit the influence of possible learning effects and position effects the remote control order (i.e. the order of the modalities in the study), task order, age group and gender were counterbalanced.

Eight participants aged from 21 to 53 years were recruited for the study. All participants watch TV on a regular basis, had previous experiences with touch-screens and had already used position sensors in smart phones. Six participants also indicated familiarity with voice input. Observation was carried out using network-cameras and microphones. The user interface prototype was running on a small form factor computer, providing the UI and TV content in Full HD resolution on a flat screen TV set.

After a short pre-interview, every participant performed both experimental conditions – the basic accelerometer-based gesture interaction condition (BC) with traditional buttons for ‘OK’ and ‘Back’, and the enhanced condition (EC) using pressure for ‘OK’ and puff for ‘Back’ functionality. For both remote controls (experimental conditions) participants had a short exploration phase followed by short questions on the first impression of system and interaction. Then participants performed three tasks where we measured task time, task success and task rating. Usability problems and errors were noted by the experimenter. After each remote control participants were asked to complete SUS and AttrakDiff questionnaire, followed by a short intermediate interviews and a final interview at the end of the study. The tasks for the study were chosen from a pool of IPTV tasks representing typical tasks for these systems (e.g. searching EPG for broadcasts, playing VOD movies, finding music in a library). Help was prepared if participants were stuck for a certain time period.

## 6 Results

**Usability Problems.** Most errors in interaction were related to precision problems with the position-sensitive navigation, including overshooting and problems finding the neutral position of the remote control. Participants were overlooking menu items in the UI and had problems with speed or accuracy for both types of interaction. On average, the enhanced condition performed faster for tasks with the EPG, while the base condition performed faster for the tasks with the music library, although no statistically significant differences was observed. Also completion rates for all tasks were investigated for differences between the enhanced condition and the base condition. Task completion overall was 73 %, but no significant difference in task completion for the two experimental conditions were observed. Furthermore, task times and ratings of perceived difficulty were investigated, but no statistical difference for the two conditions were observed. Tasks overall were perceived as difficult to complete.

**Standard Usability Score (SUS).** The novel interaction concept with the two tested conditions was assigned mediocre usability with room for improvement. The average rating for the enhanced condition was 52.8, while the base condition scored 50.3. There was no significant difference in SUS scores for study condition:  $t(14) = 0.22$ ,  $p = 0.83$ , nor for the age group:  $t(14) = 0.83$ ,  $p = 0.42$ , or the position of the remote control in the study (first or second position):  $t(14) = 0.22$ ,  $p = 0.83$ .

**User Experience – AttrakDiff questionnaire.** The enhanced condition with gesture, voice and pressure achieved higher ratings for both HQ-I (identification) as well as HQ-S (stimulation), while the ratings for the pragmatic quality (PQ) were quite low, which

is in line with findings from the SUS questionnaire and previous findings, e.g. for Touch-enabled remote controls [10]. The difference for PQ is not statistically significant ( $t(14) = 0.02$ ,  $p = 0.98$ ). Although in terms of attractiveness both remote controls achieved an above-average rating (base condition: mean = 0.93, SD = 1.66; enhanced condition: mean 1.00, SD = 1.60), no statistically significant difference was observed ( $t(14) = -0.09$ ,  $p = 0.93$ ). The overall hedonic quality of the two remote controls, combining the concepts of hedonic quality – stimulation and hedonic quality – identification, was rated high, with means between 1.33 and 1.55 on a scale from -3 to +3 for both remote controls. There was no significant difference in scores for hedonic quality ( $t(14) = -0.40$ ,  $p = 0.69$ ). Further statistical analysis took into account the gender and age group of the participants. While gender did not show significant differences in any AttrakDiff dimensions, the age group did. All but the stimulation dimension showed a significantly better rating by the younger age group (PQ:  $t(8.09) = 3.35$ ,  $p = 0.01$ ; HQ-I:  $t(14) = 3.13$ ,  $p = 0.01$ , ATT:  $t(14) = 2.96$ ,  $p = 0.01$ , HQ:  $t(14) = 2.72$ ,  $p = 0.02$ ), while the stimulation dimension only showed a trend ( $t(14) = 2.09$ ,  $p = 0.06$ ).

**First Contact and Retrospective UX Insights.** After the free exploration phase, the participants were asked for their assessment of the interaction technology in conjunction with the UI after this short period. Results indicate that traditional TV systems are still preferred by the participants of the study. Although participants were unsure if they would find everything in the user interface and would not prefer it over their traditional TV systems, the results indicate that using the user interface with these remote controls is rather fun, as both of the remote controls were rated above average. After having completed the various tasks and filling in the questionnaires for each of the remote controls, the participants answered several questions in a short intermediate interview, asking for their impressions on the remote control, how natural it felt and how precise it was perceived. 47 % of the participants overall stated that the navigation felt natural for them (33 % of answers of basic condition and 62.5 % of answers for enhanced condition). Regarding the precision, 75 % of participants ( $N = 8$ ) rated the precision of the enhanced condition as precise or rather precise, while the basic condition was rated as precise or rather precise only by 25 % of participants.

Qualitative data from the interviews confirmed the other findings of the study. Participants had problems with and criticized the interaction modalities, while at the same time liking it and having fun. Especially the enhanced condition had many positive comments about the input modalities, although also problems and the need for improvements were reported. In the final interview, when asked about preferences for one of the remote controls, participants preferred the base condition remote control. In terms of usability people preferred to have buttons and the associated haptic feedback. Seven participants evaluated the base condition remote control as more reliable, while one person stated the remotes to be equally reliable. The main reason stated for this assessment was the buttons on the base condition remote control for six of the users.

## 6.1 Limitations

In this paper, we investigated a prototypical multimodal interaction mechanism combining non-traditional input modalities for an UI that represents information in the

form of soap bubbles. However, several limitations should be noted prior to the discussion of results. The rather low number of participants might have biased the results of the statistic analysis, but allowed us to quickly gather insights on usability and user experience issues as well as qualitative insights for further development. Another possible limitation is that the novel interaction modalities and novel type of UI might have had an impact on the hedonic user experience ratings, as novelty is a part of this dimension. One possible further limitation is that we did not have a baseline condition with a standard remote control, due to the fact that the UI was not designed to be controlled with standard input modalities. To overcome this limitation, we have used standardized and validated instruments like SUS and AttrakDiff that allow for comparison across systems and with previous work in the field [10].

## 7 Discussion and Conclusion

Our research goal and motivation was to investigate the usage of non-traditional input modalities for TV content, especially regarding implications of gesture input on user experience and usability scores. The perceived usability of the participants when interacting with the novel user interface with the non-traditional interaction modalities was clearly below average. These results are supported by below-average scores for the pragmatic quality dimension of the AttrakDiff questionnaire, as well as the comments of the participants in the interviews, and are in line with findings of Vatavu [11] on the performance of gestural techniques, as well as other work in the field on non-traditional interaction modalities [10]. This clearly indicates that the interaction modalities as well as the UI need further improved in terms of usability. The high perceived task difficulty for standard TV command and control tasks additionally indicates that the usability of the system and interaction technology needs further investigation and improvements in the future.

Suggestions for improvement of the prototypes mostly addressed interaction characteristics of the modalities: the position sensitive navigation, including the precision, the reaction to the user movements, the directness and the speed of deflection.

Concerning the user experience, our research hypotheses were that through using a playful approach with the multimodal input that is also used by e.g. gaming devices a positive impact on the user experience should be observed. Ratings in hedonic quality and attractiveness were high, indicating that although the usability, reflected in the pragmatic quality dimension, was below average, the novelty of the input modality and the user interface positively influenced UX.

Interestingly, the younger group rated the AttrakDiff questionnaire significantly better than the older age group. A touch-enabled remote control on a standard IPTV UI [10] scored similar to the study conditions presented in this paper, though, which might be interpreted that modalities beyond button press might improve UX, but not necessarily usability of interfaces in a standard TV command and control context.

The study replicated previous findings [10] that the perceived user experience and perceived fun of an interaction is independent of its perceived usability, although repeated testing might be needed in order to verify the findings and account for the

limitations of this exploratory study as described in the previous section. Evaluation in situ at people's homes for extended time periods might pose a good opportunity to gain more sophisticated insights why people accept or reject new modalities and further increase the reliability of the findings of this exploratory study in the future.

Nevertheless, the findings of the present study suggest that when designing for new interaction modalities and user interfaces for the IPTV domain, the key to increase user acceptance and adoption into the users' daily lives is not only by providing a high user experience in terms of attractiveness and hedonic quality, but also to ensure a high level of usability, especially for traditional command and control tasks. Thus, a take-away of the exploratory study presented in this paper is that although new interaction modalities' user experience benefits might stand out at first sight, ensuring to maintain a high level of usability that equals or exceeds the standard button interaction that users' are used to remains a crucial point when introducing new interaction modalities or concepts into the users' living rooms.

## References

1. Bailly, G., Vo, D.-B., Lecolinet, E., Guiard, Y.: Gesture-aware remote controls: guidelines and interaction techniques. In: Proceedings of the ICME 2011, pp. 263–270 (2011)
2. Bernhaupt, R., Pirker, M., Desnos, A.: The bubble user interface: a tangible representation of information to enhance the user experience in IPTV systems. In: Proceedings of the DIS Companion 2014, pp. 85–88. ACM, New York
3. von Hardenberg, C., Bérard, F.: Bare-hand human-computer interaction. In: Proceedings of the PUI 2001, vol. 15, pp. 1–8. ACM (2001)
4. Halvey, M., Hannah, D., Wilson, G., Brewster, S.A.: Investigating gesture and pressure interaction with a 3D display. In: Schoeffmann, K., Merialdo, B., Hauptmann, A.G., Ngo, C.-W., Andreopoulos, Y., Breiteneder, C. (eds.) MMM 2012. LNCS, vol. 7131, pp. 395–405. Springer, Heidelberg (2012)
5. Hess, J., Wan, L., Pipek, V., Kuestermann, G.: Using paper and pen to control home-IT: lessons learned by hands-on experience. In: Proceedings of the EuroITV 2011, pp. 203–212 (2011)
6. Hoggan, E., Trendafilov, D., Ahmaniemi, T., Raisamo, R.: Squeeze vs. tilt: a comparative study using continuous tactile feedback. In: Extended Abstract CHI 2011. ACM (2011)
7. Kela, J., Korpijo, P., Marvi, J., Kallio, S., Savino, G., Jozzo, L., Di Marca, K.: Accelerometer-based gesture control for a design environment. Pers. Ubiquit. Comput. **10**(5), 285–299 (2006)
8. Patel, S.N., Abowd, G.D.: Blui: low-cost localized blowable user interfaces. In: Proceedings of the UIST 2007 (2007)
9. Pelling, C., Sko, T., Gardner, H.J.: Be careful how you point that thing: Wiimote aiming for large displays. In: Proceedings of the OZCHI 2009, pp. 397–400 (2009)
10. Pirker, M., Bernhaupt, R., Mirlacher, T.: Investigating usability and user experience as possible entry barriers for touch interaction in the living room. In: Proceedings of the EuroITV 2010, pp. 145–154 (2010)
11. Vatavu, R-D.: There's a world outside your TV: exploring interactions beyond the physical TV screen. In: Proceedings of the EuroITV 2013, pp. 143–152. ACM, New York (2013)

# “I Agree”: The Effects of Embedding Terms of Service Key Points in Online User Registration Form

Matjaž Kljun<sup>(✉)</sup>, Jernej Vičič, Klen Čopič Pucihar, and Branko Kavšek

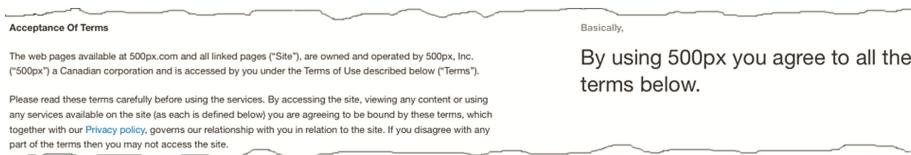
Faculty of Mathematics, Natural Sciences and Information Technologies, University of Primorska, Glagoljaška 8, Koper, Slovenia  
`{matjaz.kljun, jernej.vicic, branko.kavsek}@upr.si,`  
`klen.copic@famnit.upr.si`

**Abstract.** Terms of service (ToS) are becoming an ubiquitous part of online account creation. There is a general understanding that users rarely read them and do not particularly care about binding themselves into legally enforceable contracts with online service providers. Some services are trying to change this trend with presenting ToS section as key points on a ToS dedicated page. However, little is known how would such presentation of key points affect the continuation of user registration at the time of account creation. This paper provides an exploratory study in this area. We have offered users to participate in a draft for a prize in exchange for their names and email addresses. For this purpose we have created three registration forms: a standard form with ToS hiding behind a hyperlink and two with ToS key points presented at the time of account creation with different engagement requirements. Initial results suggest that ToS key points presented just as a list at the time of account creation is no more engaging than a form with ToS hidden behind a link. More text even made several users to complete the registration quicker than the users with the standard form. Moreover, different designs of the ToS key points list requiring different user engagement affect the interaction and reading of ToS key points, but the actual time spent on ToS is very low.

**Keywords:** Terms of service · Terms and condition · Privacy policy

## 1 Introduction

It is a common belief that Terms of Service (ToS) are written in a complicated legalese and that most users do not want to spend time on lengthy text just to create an online account. There is an emerging trend of summarising ToS and make them friendlier to read [9]. For example 500px and Pinterest have such ToS, as see on Fig. 1. In the UK it is even mandatory for certain financial services to provide a document containing “key facts” of the terms and the former Financial Conduct Authority (FCA) published “good practice” examples [10]. There is even a crowdsourced service called “Terms of Service; Didn’t Read” (ToS; DR) (<https://tosdr.org/>) that provides summaries of ToS of other web services.



**Fig. 1.** An excerpt of 500px ToS showing a summarised version of a ToS section on the left.

There are different variations of how ToS are presented to users on the web during the registration [2]. Some sites use a “clickwrap” agreement where the user clicks on “I Agree” button after seeing or scrolling through ToS. However, more commonly sites use a so-called “browsewrap” agreement where the terms are buried somewhere on the site and not showed directly to the user. Checkboxes are used to indicate that one has seen and read the terms even if these have never been shown. While the “clickwrap” forces users to actively engage with terms (even if just scrolling through), “browsewrap” solutions establish a passive engagement [8].

There is a simple reason why “browsewrap” is more common. When designing online forms the designers must make them simple to attract users and lengthy ToS do not contribute to simplicity. Even if providing ToS key points, these are not shown at the time of creating an account, but are just available on the ToS dedicated page.

To the best of our knowledge there is no research into how different user engagement with ToS at the time of account creation affects the users’ registration, hence, this paper provides an exploratory study into this area. We have created and compared three different registration forms to measure users’ engagement with ToS: (i) a “browsewrap” version, (ii) a version containing ToS key points as a bulleted list, and (iii) a variant of the later with one checkbox in front of each summarised sentence.

## 2 Related Work

There is plenty of anecdotal evidence that many users read or skim through ToS when large sums of money are involved (e.g. buying a house), medical treatment is in question (e.g. before operation), and in other similar circumstances [1]. However, on the web many users do not pay much attention when agreeing to terms, as repercussions of one’s actions commonly do not drastically affect one’s life. According to Ofcom an average internet user in the UK visited around 80 unique domains in January 2012 [7]. Reading ToS on all these domains (10 min for a ToS of an average length of 2500 words according to [6]) would take on average 13 h. It has to be acknowledged that people probably visit mostly the same domains every month and the number of new domains visited each month is lower than 80. Nevertheless, one can spend a significant amount of time reading (or just skimming through) ToS of all newly visited domains.

HCI community has been for long warning that the complexity of privacy policies, terms and conditions hinders their readability and creates one of the key usability problems of website design [3, 5, 6]. The calls for transparency, better visualisation and readability have come from academia [5], industry [9], non-government organisations [2], and from government agencies [10]. It has been already observed that a simplified

ToS can have an impact on service selectiveness. Researchers have for example shown that on a list of results of a custom made search engine consumers tend to click on the results whose ToS are ranked as more accessible and readable [8], and that users install mobile phone apps of which ToS privacy invasion visualisation is scored lower [4]. However, such visual presentations are not engaging as users see just a visual indication of how much a particular ToS invades or might invade their privacy. Moreover, a metric involved in calculating such scores could be exploited, as they are not provided by actual service owners.

In contrast with such visualisations we simply incorporated ToS key points on a registration form and tried not to affect users' inclination towards a service (e.g. not showing how much each key point invades the privacy). Such approach can maintain a certain level of simplicity of a "browserwrap" approach while still engage users with ToS, which is a characteristic of "clickwrap" forms. For this purpose we have built a web page offering users the participation in a draft for tickets for a concert in exchange for their name, surname and email. ToS were presented in three different ways as outlined in the next section.

### 3 Method

For the purpose of this study we have created three different "sign up" forms: (i) a common "browserwrap" version, (ii) a version containing a list of ToS key points, and (iii) a variant of the latter with one checkbox in front of each item on a list. Specifically, we have taken ToS Google is using for a myriad of their services (<http://www.google.com/intl/en/policies/terms/>). We have chosen this specific ToS because it is one of the mostly "agreed on" ToS currently online and it is translated in a large selection of languages (including Slovenian we have used in this study). In addition, "Terms of Service; Didn't Read" (ToS;DR) provides summarised key points of Google's ToS that we have used (retrieved on 10<sup>th</sup> of November 2014) in our study and contains 3 positive, and 6 slightly negative (in a sense they somehow invade users' privacy) key points (scale: positive, neutral, slightly negative, negative). ToS;DR classifies Google's terms as C on a scale from A to E. This means that their ToS are neither "good" nor "bad" or as they put it "*The terms of service are okay but some issues need your consideration*". We have slightly changed the terms by omitting the word Google or replaced it with the name of our institution, and added an additional section covering the research purposes and anonymisation of all data. The key points of the ToS used are listed as such:

- We keep your identifiable user information for an undefined period of time.
- We can use your content for all our existing and future services.
- This service tracks you on other websites.
- We can share your personal information with other parties.
- We may stop providing services to you at any time.
- We enable you to get your information out when a service is discontinued.
- We post a notice of ToS changes, with a 14-day ultimatum.
- We keep the rights on your content when you stop using the service.

- Partial archives of our terms are available.
- Your information will be anonymised if used for research purposes.

All three designs are visible on Fig. 2 (for the purpose of this paper the forms have been translated into English). Form 1 (left) is a common “browserwrap” implementation. Form 2 (centre) contains a list of summarised key points of ToS with a possibility to see the expanded related section of the ToS by clicking on the icon by each summarised sentence (one such expanded “key point” can be seen on the form). Form 3 (right) is a variant of the second with the addition of a checkbox in front of each summarised key point. Checking a checkbox results in highlighted key phrases of a particular sentence. In all three forms the link to whole ToS is placed above the “I agree”. Clicking on it results in opened ToS at the bottom of the form as visible on the Form 1 (the length is cut to the height of other two forms on Fig. 2). By clicking on “I agree” each participant received an email to confirm its authenticity. In the received email the link to ToS was again provided for users to visit. By clicking on “I disagree” they were presented with a form asking them why.

**Win a free ticket**

mesecneZUG VSEŠA

Fill out the form and win a concert ticket for Zabljajena generacija on 17. 12. 2014 (22:00) in Planet TUŠ in Koper.

Name:  Winet me

Surname:  Winet priimek

E-mail:  Winet vse elektronski naslov

I am familiar with the terms of service.

I agree  I disagree

You can contact us about any question regarding the terms of service: [projekt@metf.uni-lj.si](mailto:projekt@metf.uni-lj.si).

**Terms of Service**

Welcome!

Thanks for using our products and services (“Services”). The Services are provided by the Faculty of Mathematics, Natural and Information Sciences, Institute Dvignitev 8, 3900, Koper, Slovenia.

**Win a free ticket**

mesecneZUG VSEŠA

Fill out the form and win a concert ticket for Zabljajena generacija on 17. 12. 2014 (22:00) in Planet TUŠ in Koper.

Name:  Winet me

Surname:  Winet priimek

E-mail:  Winet vse elektronski naslov

Summary of terms:

- We keep your identifiable user information for an undefined period of time.
- We can use your content for all our existing and future services.
- This service tracks you on other websites.
- We can share your personal information with other parties.
- Google+ profil storitev
- Zadra sprememb: 14. apr. 2014 (igkeit aktivnosti razreda)
- Partial archive of our terms are available.
- Our information will be anonymised if used for research purposes.

I am familiar with the whole terms of service.

agree  I disagree

You can contact us about any question regarding the terms of service: [projekt@metf.uni-lj.si](mailto:projekt@metf.uni-lj.si).

**Win a free ticket**

mesecneZUG VSEŠA

Fill out the form and win a concert ticket for Zabljajena generacija on 17. 12. 2014 (22:00) in Planet TUŠ in Koper.

Name:  Winet me

Surname:  Winet priimek

E-mail:  Winet vse elektronski naslov

Summary of terms:

- I read the below.
- We keep your identifiable user information for an undefined period of time.
- We can use your content for all our existing and future services.
- This service tracks you on other websites.
- We can share your personal information with other parties.
- We may stop providing services to you at any time.
- We may change the terms of service at any time. The new terms of service is described.
- We can terminate the right to your content, when you stop using it.
- We post notices of changes, with a 14-day ultimatum.
- Partial archive of our terms are available.
- Our information will be anonymised if used for research purposes.

I read the above.

I am familiar with the whole terms of service.

agree  I disagree

You can contact us about any question regarding the terms of service: [projekt@metf.uni-lj.si](mailto:projekt@metf.uni-lj.si).

**Fig. 2.** Three different designs of a “sign up” form. Form 1 (left) is a common “browserwrap” form with expanded ToS on the bottom (achieved by clicking on a ToS link above the buttons). Form 2 (centre) is a form with ToS key points; each key point can be expanded by clicking on the info icon in front of it. Form 3 (right) is a variant of the later with a checkbox in front of each ToS key point; when checked the key phrases of a particular summary were highlighted.

### 3.1 Dissemination and Participants

We chose a Christmas concert for students organised by a shopping centre. With their consent we created the above-described forms. The site was accessible from 28th of November 2014 to 17th of December 2014 for a total of 20 days. The QR code leading to the web page was positioned in a corner of posters advertising the concert two weeks before the event happened. However, the code was not accompanied by any text and it occupied just a small part of the poster. The same code and a link to the form were put on a few web sites advertising events one week before the event.

The target population were students who for exchange of their name, surname and email address entered a draft for 30 free concert tickets. The prize itself was not high as the tickets could be bought for just 3€. Our belief is that the higher the price (or the added value of a service) the higher is users' willingness to sacrifice their online privacy. With the low price users were thus not inclined to continue with registration if they perceived ToS as too invading or the process too lengthy.

When users visited the site one of the three forms was randomly showed. We logged the time spent on the form, the mouse movements and clicks using the Clickheat software to capture users' engagement with the site and in particular with ToS. At the same time we used Apache logs filtered and summarised by AWStats.

## 4 Results

During the period of 20 days, 340 unique users (excluding spiders, and other foreign IP addresses) visited the web site. Each was presented with exactly one of the three forms (see Fig. 2 and the description in Sect. 3): Form 1 (F1) was a “browserwrap”, Form 2 (F2) included TOS key points, and Form 3 (F3) had key points highlighted when checked as read. The number of times each form was visited is presented in Table 1.

**Table 1.** Number of visits per form and numbers of how many visitors completed each form.

	Form 1	Form 2	Form 3
Number of times shown by the random algorithm	166	92	82
Number of times each form was completed	22	8	7
Percentage of completed forms in relation to visits	13.3 %	8.7 %	8.5 %

The last line in the Table 1 presents percentages of the forms completed. The numbers show a substantial dropout for all three. However, the dropout for F2 and F3 (forms with summarised ToS) was even higher. This can be interpreted in two different ways: either users were “scared away” by the length of the text they were presented with and the prize was not worth the effort, or they had actually read the key points of terms and were not willing to give up their data for a small prize. Even more interesting is that the percentage of people who completed F2 and F3 is similar even though users completing F3 had to make 10 more clicks. The fact that there were more clicks increased ToS engagement of F3 if compared to F2. However, the average time spent on each form is not high (see the left column in Table 2): 37 s for F1, 36 s for F2, and 59 s for F3. This suggests that visitors visiting F1 and F2 did not engage much with ToS and did not spend much time reading. Moreover, more users completed F2 under 30 s than F1 as if more text would make them hurry to complete the form.

The higher average time spent on F3 (see Table 2) is due to 10 additional clicks on checkboxes, which F2 did not require. This contributed to 2 additional second for each checkbox checked for F3. These 2 s are also enough to skim the associated ToS key point and check it as you read. If looking at individual time frames in **Error! Reference source not found.**, visitors who spent between one and two minutes on the form had

enough time to skim through the text presented on F2 and F3. Of all visitors only 7 (2 %) actually opened the stand-alone ToS page linked from the confirmation email.

**Table 2.** Times spent on each form and percentage of people completed each form in different time frames

	Average time in seconds to complete	% of those who spent			
		< 30s	30s – 60s	60s – 120s	> 120s
<b>Form 1</b>	39s	27.3 %	54.5 %	13.6 %	4.5 %
<b>Form 2</b>	37s	50.0 %	37.5 %	12.5 %	0.0 %
<b>Form 3</b>	59s	14.3 %	28.6 %	57.1 %	0.0 %

The interaction with the forms can be seen on Fig. 3. Clicks are shown on a blue-green-yellow scale from low (blue) to high (yellow) number of clicks. On F1 and F2 visitors often clicked on the link (bordered purple) above the “I Dis/Agree” buttons that showed the ToS below the form. Interestingly, on F2 visitors rarely clicked on information links by each ToS key point. By clicking on these information icons before each key point the related ToS section was revealed. However, the F3 shows the opposite. Visitors were forced to interact with ToS key points and often clicked on the information icons by the checkboxes to reveal the related ToS section. This is particularly visible by the top key points (bordered orange), while the frequency of clicks waters down with lower key points. One possible explanation is that they were already overwhelmed by clicking. Nevertheless, roughly 60 % of those who completed F3 and 13 % of those who completed F1 and F2 (or 24 % all together) skimmed through or read ToS to some extent.



**Fig. 3.** Interaction with the forms. The number of clicks in each particular spot is visualised on a blue-green-yellow scale from low (blue) to high (yellow) number of clicks (Color figure online).

## 5 Discussion

This preliminary study presents the exploratory results into how users interact and engage with summarized ToS into key points at the time of an online account registration. The initial data suggests that summarised ToS is no more engaging than a simple “browserwrap” version that has ToS hidden from the user. Different designs just contribute to different ways of interaction and engagement. Nevertheless, a closer look at the interaction shows that 24 % of users who registered for the service at least skimmed through ToS (60 % of those completed F3 with most interaction involved). Higher engagement can also lead to higher enforceability in the legal terms [2].

We admit that this study has several limitations. Firstly, we had some technical glitches that need to be dealt with for the follow up study. One is the randomisation of the form distribution. We used PHP’s random function, which resulted in displaying F1 nearly twice as much as the other two forms. Simply alternating forms would ensure sufficient randomisation as visitors are randomly visiting the site. The Clickheat software was also not recording all clicks, which is due to poor documentation and using default values. However, these glitches have not affected the study process.

In addition to technical, some methodology limitations were present as well. For this preliminary study we have not conducted any formal interviews. However, we had some informal interviews with people who received the tickets. Those who were presented with forms featuring ToS key points revealed that some looked slightly intimidating. This suggests that the selection of key points from ToS;DR might have contributed to a higher dropout as the forms with ToS key points showed nearly a third as much dropout than the form without. Nevertheless, even the form with no ToS key points experienced 86 % dropout. In the future runs of this experiment more interviews need to be conducted to understand this phenomenon in details. The mere presence of 10 sentences (and checkboxes on one form) might be behind some dropout as well since the prize in question was not particularly attractive. Moreover, when presented with the text and no actual engagement (Form 2), users who completed the registration tended to complete it quicker compared to the form with no text (Form 1).

Another limitation is the target population for the event chosen – namely students. Nevertheless, this group of users is very active on the internet and presents a diverse range of students from different study programs including humanities, language studies, education studies, management, mathematics, kinesiology, biology and computer science. We can also assume that students at our university come from diverse social backgrounds and diverse geographical locations, and are aged between 18 and 24. We can assume that generalisation over this age group is reasonable while the results can be thus hardly generalised over whole population.

## 6 Conclusion

ToS key points are a new alternative to lengthy legalise language. However, the web sites are listing them only on ToS dedicated pages. In this paper we provide an initial view into how users interact and engage with ToS key points at the time of account

creation for an online service. Whilst noting that these results are of preliminary nature, they clearly show trends worth presenting. The main findings reveal a high dropout between users who visited and those who completed the form (as high as 91.5 %) with forms featuring summarised key points having an even greater dropout. The average time spent on each form was also low. Nevertheless, the data suggest that different designs contribute to different engagement with ToS and that 13 % of those less engaged to 60 % of those most engaged at least skimmed through the ToS text.

The presented study raises some questions such as why so many visitors had left the page. While we have speculated about the possible reasons (too much to read, low priced reward, intimidating ToS key points) we would have to back up this data. We are planning to address these issues in the future runs of the study with qualitative (interviews) and quantitative approach (attracting a larger and more diverse population). The future work includes also the elimination of technical problems of the experiment.

## References

1. AskReddit (banjomatic): People who read through the Terms & Conditions, did you ever find anything to stop you from accepting. [http://www.reddit.com/r/AskReddit/comments/2qs3g2/people\\_who\\_read\\_through\\_the\\_terms\\_conditions\\_did/](http://www.reddit.com/r/AskReddit/comments/2qs3g2/people_who_read_through_the_terms_conditions_did/)
2. Bayley, E.: The Clicks That Bind: Ways Users ‘Agree’ to Online Terms of Service. Electronic Frontier Foundation (2009)
3. Fiesler, C., Bruckman, A.: Copyright terms in online creative communities. In: CHI EA 2014, pp. 2551–2556. ACM, New York (2014)
4. Liccardi, I., et al.: No technical understanding required: Helping users make informed choices about access to their personal data. In: MOBIQUITOUS 2014, pp. 140–150. ICST, Brussels, Belgium (2014)
5. Luger, E., et al.: Consent for all: revealing the hidden complexity of terms and conditions. In: CHI 2013, p. 2687. ACM, New York (2013)
6. McDonald, A.M., Cranor, L.F.: The cost of reading privacy policies. *A J. Law Policy Inf. Soc.* **543**(4), 1–22 (2008)
7. Ofcom: On average, internet users are visiting fewer domains. <http://stakeholders.ofcom.org.uk/market-data-research/market-data/communications-market-reports/cmr12/internet-web/uk-4.30>
8. Tsai, J.Y., et al.: The effect of online privacy information on purchasing behavior: An experimental study. *Inf. Syst. Res.* **22**(2), 254–268 (2011)
9. UXmovement (anthony): Why Every Terms of Service Page Needs Summaries. <http://uxmovement.com/content/why-every-terms-of-service-page-needs-summaries/>
10. Waters, D.: Good and poor practices in Key Features Documents. The Financial Conduct Authority (2007)

# Automatic Privacy Classification of Personal Photos

Daniel Buschek<sup>1()</sup>, Moritz Bader<sup>1</sup>, Emanuel von Zezschwitz<sup>1</sup>,  
and Alexander De Luca<sup>1,2</sup>

<sup>1</sup> Media Informatics Group, University of Munich (LMU), Munich, Germany

{daniel.buschek, emanuel.von.zezschwitz,

alexander.de.luca}@ifi.lmu.de, moritz.bader@googlemail.com

<sup>2</sup> DFKI GmbH, Saarbrücken, Germany

**Abstract.** Tagging photos with privacy-related labels, such as “myself”, “friends” or “public”, allows users to selectively display pictures appropriate in the current situation (e.g. on the bus) or for specific groups (e.g. in a social network). However, manual labelling is time-consuming or not feasible for large collections. Therefore, we present an approach to automatically assign photos to privacy classes. We further demonstrate a study method to gather relevant image data without violating participants’ privacy. In a field study with 16 participants, each user assigned 150 personal photos to self-defined privacy classes. Based on this data, we show that a machine learning approach extracting easily available metadata and visual features can assign photos to user-defined privacy classes with a mean accuracy of 79.38 %.

**Keywords:** Photos · Privacy · Classification · Images · Metadata

## 1 Introduction

Browsing a personal photo gallery in the presence of others can unintentionally reveal private content and therefore violate the user’s privacy. Moreover, users may want to share their photos online on social platforms like Flickr or Facebook. To support users in revealing their photos only to intended audiences, applications need to become more privacy-aware, as recommended in related research [1, 10, 11].

To achieve this, applications should be informed about the user’s intended privacy setting, ideally for each photo individually [1]. Unfortunately, manually sorting photos with respect to privacy concerns and different audiences is time-consuming or not feasible at all for large collections.

We present an approach to automatically assign personal photos to user-defined privacy classes. For example, users might define classes related to specific places, activities, events or audiences. We employ a machine learning approach to infer each photo’s intended privacy class from metadata (e.g. timestamp, GPS, ISO-speed) and from visual features (e.g. based on colours, edges, occurrences of faces). Our insights enable more privacy-aware photo applications and thus support users in sharing their photos only with their intended audiences. Our contribution is twofold:

- *We describe and evaluate metadata and visual features with respect to users' own privacy classification.* In a user study ( $N = 16$ ), participants sorted personal photos into three self-defined privacy categories. We show that a Random Forest classifier matches these user-defined categories with 79.38 % accuracy.
- *We present a privacy-respecting study method to gather relevant image data from personal photos.* We implemented a tool for users to extract image data themselves at home. Hence, they never had to reveal their photos.

## 2 Related Work

Ahern et al. [1] investigated privacy decisions and considerations in mobile, online photo sharing. They examined photo tags and privacy settings with a Flickr app, showing that tags related to persons and locations had the highest ratio of private photos. They derived a taxonomy of privacy considerations from their data analyses and interviews. This revealed that users' main concerns were related to influencing their own online identity and those of others. The authors concluded that applications should help users to prevent privacy-related mistakes when sharing photos online. This supports the motivation for our work. Furthermore, the results from our feature analysis match their findings regarding the influence of persons and locations on users' privacy settings.

Zerr et al. [11] targeted privacy-aware image search with a Support Vector Machine (SVM) classifier trained on publicly available photos from Flickr. They used five visual features: faces, hue histogram, edge-direction coherence vector [7], SIFT features [5], and average brightness and sharpness. Their search method achieved 0.74 F<sub>1</sub>-Score in retrieving private versus public photos, and 0.80 when combined with title and tags as textual features. They did not report classification accuracy.

They also proposed an alert system to warn users when uploading potentially sensitive photos [10]. However, in both projects [10, 11], the “private” photos had been published on the web by Flickr users. Hence, these pictures were not considered private by their owners. They were later tagged as private by others via a community game. In contrast, we classify personal photos not shared on the web. We employ similar visual features, but do not use textual annotations. We further include metadata features, such as timestamps and GPS locations.

Klemperer et al. [3] repurposed existing organisational image tags for access control in photo sharing. They concluded that: “It may be possible to additionally aid users with [...] automated tag generation”. This motivates our idea of adding new privacy tags with automatic privacy classification into user-defined classes. Klemperer et al. employed decision trees to generate access rules from tags, while we employ them to add tags from metadata and visual features. Their participants tagged photos in the lab, whereas ours sorted their photos at home. Since we aim to tag photos automatically, we are not interested in observing users, but rather in including photos, which users might not like to bring to a lab study. We present an evaluation concept to include such photos while respecting participants' privacy.

### 3 Approach

We first describe a threat model and explain how a system using our photo classification method protects the user’s privacy in related scenarios. Thereafter, we describe our photo classification system in more detail.

#### 3.1 Threat Model

When users are browsing through a photo gallery on their mobile device in the presence of others, such as friends, family members, or unknown passengers on a bus, these bystanders could (un)intentionally catch a look at the pictures on the screen, thus possibly violating the users’ privacy. The user might also *want* to present some pictures to others, but the gallery could then also reveal private ones while browsing. In another scenario, the user uploads pictures to a social network, but only wants to share certain pictures with certain groups of people.

To avoid revealing private pictures to unwanted eyes, we imagine applications to allow users to create privacy classes and to switch between them, for example “myself”, “colleagues”, “friends”. Only pictures assigned to the current setting are then displayed. However, this requires the user to assign a privacy class to each picture - a potentially very tedious and time consuming task.

The system proposed in this paper addresses this issue by automatically assigning one of three user-defined privacy classes to each new picture.

#### 3.2 Photo Classification System

We employ a machine learning approach to assign photos to privacy classes. Our method comprises three steps: First, we define and extract relevant features from metadata and the visual pixel data itself. Second, we train a classifier on these features, extracted from a set of training images. Finally, the trained classifier can assign new photos to privacy classes.

**Features.** We extracted two types of features: metadata (location, time, shot details) and visual features (faces, colours, edges). All features are described in Table 1, which also provides references to related work for in-depth descriptions.

The choice of examined features was heuristically guided by expectations regarding possible indicators for privacy. For example, certain locations and timeframes could be related to certain activities in the user’s life, such as a holiday, sports training, nightlife, and so on. Moreover, faces reveal the presence of people in a photo; a single face might indicate a more private setting than a group of many faces. Additionally, long straight edges indicate man-made structures (e.g. indoors, in a city), while scenes in nature feature many short incoherent edges [2, 7]. Our best feature set, derived from analysis of our study data, indeed contains features related to location, time, edges and the number of faces.

**Classifiers.** We evaluated three common classification approaches to show that the described features can be suitably used by different methods. In particular, we evaluated: Random Forest (RF), Support Vector Machine (SVM), and Nearest Neighbour (NN). Random Forest performed best.

## 4 Field User Data Collection

We conducted a field study to collect user data and evaluate our approach: Users defined three privacy classes and manually assigned 50 personal photos to each class. They extracted features with a given application and sent us the resulting feature-file. Hence, we never saw the users' actual photos. Photos cannot be reconstructed from the described features.

**Participants:** We recruited 17 participants with an average age of 26 years (SD 9). 10 were female, 7 male. One participant was later excluded from analysis, since this user had taken all photos exclusively for this study, which renders the data artificial. Participants were compensated with a 15€ gift card for an online shop.

**Apparatus:** Users were given a simple application without a graphical interface. When executed, it extracted the relevant metadata and visual features from all photos within a provided folder and wrote them to a feature-file.

**Procedure:** Users participated remotely. We sent them the feature extraction application and study instructions: First, they defined their own privacy classes by creating a folder for each and naming it; they also added a privacy ranking (public: 1 to private: 5), for example “5\_myself”. Second, users browsed their photo collections for 50 pictures per class and copied them into the corresponding folders. Third, they ran our feature extraction application and uploaded the resulting file to our server. Users also filled in a short questionnaire. Finally, we asked them to assign 5 new photos to each of their privacy classes after a week.

## 5 Results

We used the scikit-learn library [6] for Python to implement and evaluate the proposed system. We report classification accuracy; the ratio of photos for which the automatic assignment matches the user's manual assignment. Accuracies were computed with 10-fold stratified cross-validation.

### 5.1 Feature Selection

We first evaluated classification accuracy when using each feature on its own. The classifiers' hyperparameters were optimised per feature for each user. Table 1 shows the results: Overall, time and location features performed best.

We then applied a wrapper feature subset selection approach [4] to find the best combination of features. To reduce the search space for the wrapper, we removed the least promising features - those which never appeared among the top half in at least one of three tests: single feature evaluation, ANOVA F-value-score, and feature importances with RF. We refer to the library documentation for further details and related reading [6]. Our wrapper method greedily tests feature sets with a given classifier (here we used RF, since it performed best for single features) and removes the feature for which the remaining set leads to the best classification accuracy.

**Table 1.** Single feature evaluation. For each feature, the table shows mean classification accuracy and standard deviation achieved with the three tested classifiers, when considering only this feature. Features are ranked by resulting maximum accuracy with any of the three tested classifiers. The last column shows for how many users this feature was extracted from the collected data (feature present in > 50 % of a user’s photos). Highlighted are the features comprising the best feature set, as found with a wrapper subset selection approach.

Feature	Description	RF		SVM		NN		Users
		mean	std	mean	std	mean	std	
1 unix mins	minutes from the photo’s unix timestamp	73.75	14.65	70.46	16.41	69.67	17.34	16
2 latitude	latitude from the photo’s GPS data	70.42	11.2	55.21	14.85	67.37	14.53	16
3 longitude	longitude from the photo’s GPS data	69.79	13.37	55.42	15.97	67.17	13.99	16
4 unix days	days from the photo’s unix timestamp	68.54	14.96	65.71	15.15	60.88	13.71	16
5 day-time	timestamp relative to the day from the photo’s unix timestamp	63.75	14.92	57.04	13.63	61.67	13.48	16
6 address	country code and postcode derived from the photo’s GPS data	63.38	14.67	61.50	14.44	55.75	15.6	16
7 elevation	metres above sea level from the photo’s GPS data	63.00	15.55	60.17	14.07	58.96	15.16	16
8 cal-week	calendar week of the year from the photo’s unix timestamp	62.75	13.7	60.83	14.52	56.04	12.33	16
9 hist-hue	histogram (16 bins) of the photo’s hue distribution over all pixels	59.92	10.72	56.50	10.48	55.67	11.74	16
10 month-day	day of the month from the photo’s unix timestamp	59.00	14.74	54.71	15.02	54.46	13.8	16
11 edcv	histogram (36 bins) of the angle distribution of detected edges [7]	52.92	10.19	50.37	9.95	49.17	9.69	16
12 exposure	exposure from the photo’s metadata	52.52	6.76	44.67	9.17	48.14	5.38	14
13 hist-bright	histogram (16 bins) of the photo’s brightness distribution over all pixels	52.29	9.83	46.71	9.5	49.33	8.96	16
14 hist-sat	histogram (16 bins) of the photo’s saturation distribution over all pixels	51.87	10.11	47.17	9.84	47.29	10.45	16
15 ISO-speed	ISO-speed value from the photo’s metadata	50.96	9.72	49.71	8.56	47.17	9.56	16
16 imp_hue	index of the largest bin of the hue-histogram	49.42	11.83	50.33	10.91	46.83	10.06	16
17 weekday	day of the week from the photo’s metadata	49.25	11.01	48.87	11.42	45.13	10.14	16
18 acutance-local	edge strength (Sobel operator) of detected edge pixels (Canny edge detector)	41.33	9.06	46.37	6.65	43.46	8.97	16
19 holiday	flag indicating whether the day from the photo’s metadata is a public holiday	44.10	8.29	44.26	8.75	39.28	8.2	13
20 edge-ratio	ratio of detected edge pixels (Canny edge detector) to the total number of pixels	42.58	7.22	43.38	8.3	44.17	7.23	16
21 flash	flag indicating whether the flash was used, taken from the photo’s metadata	43.43	8.39	43.62	7.89	42.43	8.17	14
22 num faces	number of faces detected in the photo [8, 9]	42.58	6.09	42.54	5.65	42.17	5.2	16
23 edcv-ratio	ratio of coherent edges to non-coherent edges [7]	42.17	10.07	42.38	8.42	41.79	10.79	16
24 focal-len.	focal length from the photo’s metadata	41.96	11.4	41.21	11.23	40.50	10.05	16
25 acutance-global	edge strength (Sobel operator) of every 9th pixel	41.71	8.8	41.96	7.23	41.13	9.05	16
26 brightness	perceived brightness $(0.299R + 0.587G + 0.114B)$ averaged over all pixels	40.54	9.48	40.12	9.67	41.17	8.61	16
27 resolution	number of pixels	41.17	11.69	40.96	11.74	39.17	9.58	16
28 f-number	f-number from the photo’s metadata	41.00	11.27	40.14	11.2	38.95	9.98	14
29 edcv-std-c	standard deviation of the angle of coherent edges [7]	39.29	5.79	39.79	6.22	40.08	7.18	16
30 orientation	flag indicating “portrait”, “landscape” or “square”	39.96	3.42	39.79	3.66	35.71	5.07	16
31 edcv-std-n-c	standard deviation of the angle of non-coherent edges [7]	38.58	7.72	38.67	5.92	37.46	6.21	16
32 model	camera model from the photo’s metadata	38.33	7.26	38.37	7.2	35.75	5.02	16

The search terminates when no further improvement can be achieved with changes to the feature set.

The best found set consists of 12 features (**highlighted** in Table 1) latitude, longitude, elevation, Unix minutes, calendar week, weekday, day of the month, important hue, ISO-speed, local acutance, number of faces, and resolution.

This set includes features from all examined dimensions: location, time, shot details, face detection, colours, and edge detection. Hence, all described dimensions were found to be relevant and complement each other for privacy classification of personal photos. The best classifier (RF) achieved 79.38 % (SD 11.00 %) classification accuracy with this optimised set.

## 5.2 Error Case Analysis

Although users were free to define any three privacy classes, not necessarily hierarchical ones, all created a hierarchy. Therefore, given our threat model, we can distinguish two types of errors: assigning an image to a less private class, or assigning it to a class of higher privacy. We can expect the first type to be more serious in most applications, since it possibly reveals private content to the wrong audience. The second type could cause manual correction efforts, but no privacy violations (assuming unambiguously hierarchical classes, as chosen by our participants). An analysis of serious mistakes showed that their ratio was close to 50 % for all classifiers (RF: 49.49 %, SVM: 49.00 %, NN 48.15 %). Hence, our classifiers were not biased towards serious classification errors.

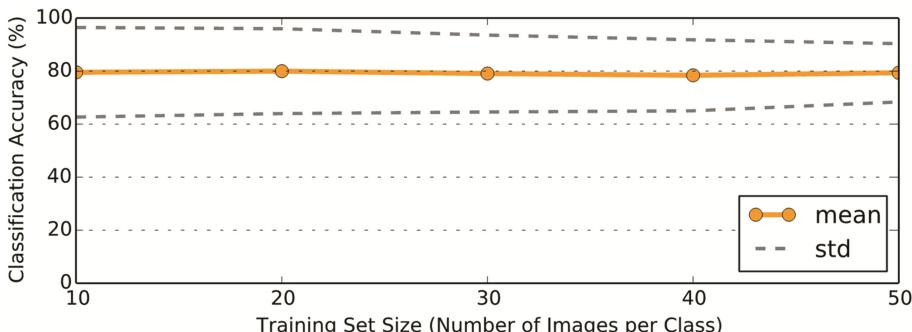
### 5.3 Variability Over Time

To test variability in user behaviour over time, we asked users to manually assign 5 new photos to each of their privacy classes after a week. For each user, we then trained our system on all 150 photos from the first session and evaluated how well it could assign the 15 new photos. We observed a mean accuracy of 55.42 % (SD 19.82 %). While still better than random guessing (33 %), this is a clear decline from the cross-validation results within one session (79.38 %).

To further investigate this, we asked users whether their 15 new photos were specifically taken for the study. Unfortunately, half of them had taken at least some new photos for the study. This explains the vast decline in accuracy, since those pictures were likely taken artificially in a short timeframe and at the same location, rendering two of our main features mostly useless. Accordingly, when only evaluating data for the other half of users, mean accuracy again increased to 67.50 % (SD 20.39 %). We explain the remaining gap to the results within sessions - 82.00 % (SD 12.83 %) for this subset of users - with changes in users' mental models regarding their self-defined privacy classes.

### 5.4 Training Set Sizes

Our system requires labelled photos for training. In practice, the user thus has to manually assign some photos to each privacy class. Hence, it is interesting to evaluate how well the system performs with fewer training images. Figure 1 shows that standard deviations increase with less training data, but mean accuracy stays consistent. In conclusion, these results show that classification with 10 manually assigned images per class is on average as good as with 50 training images.



**Fig. 1.** Classification accuracy (RF) as a function of training set size. The plot shows that mean accuracy stays consistent when using fewer manually assigned photos for training.

### 5.5 Comparison to Human Reasoning

We proposed and employed a study method which respects users' privacy. Since we thus never saw their actual photos, we asked users to provide general comments on their own manual classification procedures through a questionnaire.

We analysed these comments looking for heuristics similar to our features. Users mentioned related criteria such as: presence of people (12 of 17 users); landscape and architecture (5 users); certain places or events, like “at the beach” or during a specific holiday (7 users). These considerations support the use of face detection, edge detection, time and location features, although human reasoning is of course much more complex than what can be inferred with these features. For example, regarding pictures of people, users mentioned refined criteria such as viewing angle, facial expressions, certain types of clothing, and specific individuals. Nevertheless, users’ comments suggest that metadata and visual features can to some extent capture relevant aspects of human reasoning regarding privacy classes of personal photos.

## 6 Limitations

Extracting simple metadata and visual features, a Random Forest classifier achieved 79.38 % classification accuracy. Half of the remaining errors were assignments to less private classes. Hence, about 10 % of all automatically assigned photos could potentially violate the user’s privacy, assuming hierarchical classes as created by our users. To address this issue, the automatic assignment could be presented to the user for a final check (e.g. thumbnails grouped per class).

We observed lower accuracies for assigning photos a week after training. Changing mental models of privacy classes present a challenge to automatic classification. However, classifiers could be retrained with a few new manually assigned photos to adapt to the user’s mental model continuously. Additionally, long-term use of applications employing privacy classes might lead to more stable user perspectives on their self-defined classes. We leave this analysis to a future study. Nevertheless, our current system still clearly outperformed random guessing after a week.

## 7 Conclusion and Future Work

Privacy is an important issue when browsing personal photos on a mobile device in the presence of others, or when sharing photos online. Privacy-aware photo applications need information about the privacy class of each picture. This can lead to tedious manual assignments.

Thus, we have presented a system to automatically assign personal photos to user-defined privacy classes. In a field study with 16 participants, we showed that our machine learning approach can classify users’ photos with an average accuracy of 79.38 %, based on easily available metadata and visual features. In conclusion, our approach can enhance privacy-aware applications by automatically providing privacy classes for filtering photos. Possible application scenarios include privacy-aware gallery browsing on a mobile device and selective photo sharing in a social network.

Furthermore, we have described a study method to gather relevant image data (metadata and features) without revealing the actual pictures to the researchers. We expect this approach to be useful for other studies with personal photos.

We plan to implement our system in a mobile gallery application for a deployment “in the wild”. Data from long-term use can then be analysed to examine changes in users’ mental models over time. Further advanced computer vision techniques could be investigated to boost accuracy, for example object recognition.

## References

1. Ahern, S., Eckles, D., Good, N.: Over-exposed?: privacy patterns and considerations in online and mobile photo sharing. In: CHI 2007, pp. 357–366 (2007)
2. Kim, W., Park, J., Kim, C.: A novel method for efficient indoor-outdoor image classification. *J. Sig. Process. Syst.* **61**(3), 251–258 (2010)
3. Klemperer, P., Liang, Y., Mazurek, M., Sleeper, M., Ur, B., Bauer, L., Cranor, L. F., et al.: Tag, you can see it! using tags for access control in photo sharing. In: CHI 2012, pp. 377–386 (2012)
4. Kohavi, R., John, G.H.: Wrappers for feature subset selection. *Artif. Intell.* **97**(1–2), 273–324 (1997)
5. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **60**(2), 91–110 (2004)
6. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., et al.: Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011)
7. Vailaya, A., Jain, A., Zhang, H.J.: On image classification: City images vs. landscapes. *Pattern Recogn.* **31**(12), 1921–1935 (1998)
8. van der Walt, S., Schönberger, J.L., Nunez-Iglesias, J., Boulogne, F., et al.: Scikit-image: Image Processing in Python. *PeerJ*:e453 (2014)
9. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: CVPR 2001, pp. I–511–I–518, vol. 1 (2001)
10. Zerr, S., Siersdorfer, S., Hare, J.: PicAlert!: a system for privacy-aware image classification and retrieval. In: CIKM 2012, pp. 2710–2712 (2012)
11. Zerr, S., Siersdorfer, S., Hare, J., Demidova, E.: Privacy-aware image classification and search. In: SIGIR 2012, pp. 35–44 (2012)

# CipherCard: A Token-Based Approach Against Camera-Based Shoulder Surfing Attacks on Common Touchscreen Devices

Teddy Seyed<sup>1()</sup>, Xing-Dong Yang<sup>2</sup>, Anthony Tang<sup>1</sup>, Saul Greenberg<sup>1</sup>, Jiawei Gu<sup>3</sup>, Bin Zhu<sup>4</sup>, and Xiang Cao<sup>5</sup>

<sup>1</sup> Department of Computer Science, University of Calgary, Alberta, Canada  
`{teddy.seyed, tonyt, saul}@ucalgary.ca`

<sup>2</sup> Department of Computer Science, Dartmouth College, Hanover, NH, USA  
`xing-dong.yang@dartmouth.edu`

<sup>3</sup> Baidu Institute of Deep Learning, Beijing, China  
`gujiawei@baidu.com`

<sup>4</sup> Microsoft Research Asia, Beijing, China  
`binzhu@microsoft.com`

<sup>5</sup> Xiaoxiaoniu Creative Technologies, Beijing, China  
`xiangcao@acm.org`

**Abstract.** We present CipherCard, a physical token that defends against shoulder-surfing attacks on user authentication on capacitive touchscreen devices. When CipherCard is placed over a touchscreen’s pin-pad, it remaps a user’s touch point on the physical token to a different location on the pin-pad. It hence translates a visible user password into a different system password received by a touchscreen, but is hidden from observers as well as the user. CipherCard enhances authentication security through Two-Factor Authentication (TFA), in that both the correct user password and a specific card are needed for successful authentication. We explore the design space of CipherCard, and describe three implemented variations each with unique capabilities. Based on user feedback, we discuss the security and usability implications of CipherCard, and describe several avenues for continued exploration.

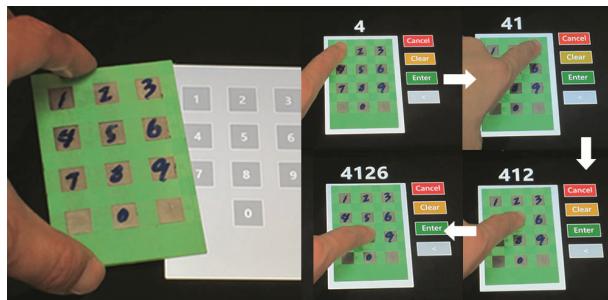
**Keywords:** Shoulder-surfing attack · Capacitive touchscreen · PIN entry · Security

## 1 Introduction

Capacitive touchscreens have become the primary input mechanism for many security related applications such as access control systems (e.g. door locks), public kiosks (e.g. ATMs, cash registers, point of sales via large screens), or mobile authentication (e.g. payment through personal mobile devices). Because user authentication through touchscreens is often carried out in public spaces, the user is susceptible to shoulder-surfing attacks: unscrupulous individuals or cameras can see the password or PIN being entered into the system [1–6]. Further exacerbating the problem, user interfaces for

touchscreens are often designed to be larger (because of the fat finger problem [7]), making it difficult to shield input from observation. As well, the lack of haptic feedback on touchscreens makes eyes-free operation difficult, which means that users cannot easily shield the display from view.

To enhance user authentication security on touchscreen devices, we present CipherCard, a physical token that protects a users' PIN entry against camera-based shoulder-surfing attacks. CipherCard (Fig. 1) is an opaque overlay that is placed atop a touchscreen's password input area (e.g., a touchscreen PIN pad), where it serves as a physical proxy for the touchscreen's original password input UI. When a user touches a button on the CipherCard, this touch point is remapped to a different button location on the touchscreen via its internal wiring, hiding the actual input location from observers or hidden cameras. CipherCard translates the input sequence ("user password") into a distinct sequence ("system password") that is received by the touchscreen. For example, Fig. 1 illustrates a user entering their user password of '1 3 5 8' into CipherCard, which is translated to the system password of '4 1 2 6'. Thus, the system password is hidden from observers or hidden cameras. A shoulder-surfing attacker may acquire the user password, but without the user's CipherCard, the attacker cannot successfully authenticate.



**Fig. 1.** CipherCard maps a touch into a different location.

CipherCard mappings can be permanent (where they are manufactured with a single translation for use on particular systems) or reconfigurable (where a user can specify the translation between user and system passwords). CipherCard allows a user to choose a set of easy-to-memorize user passwords, and use them as proxies for "strong" system passwords (e.g. PINs with random combinations). Renewing a system password can be as easy as getting a new CipherCard or reconfiguring an existing one, where the user password can remain the same. CipherCard can be designed to automatically adapt to different password input UIs (e.g. size, orientation, and layout), and vice versa.

The CipherCard authentication scheme raises many engineering, security, and usability questions that warrant long-term research. At this early stage, we focus on a thorough exploration of the design space. We implemented three prototypes: card-shaped, wallet-based, and phone-based CipherCards, each with a unique form factor and usability features. To evaluate our progress, we conducted two studies: first, a feasibility evaluation with five usability professionals, where we identify usability issues of the

three prototypes. Based on the findings, we proposed an improved design. Second, we conducted a separate workload evaluation to verify our improved design as well as the overall usability and security of CipherCard.

Our contributions in this paper are three-fold: (a) a novel CipherCard concept that allows remapping for key entries to protect PIN input against camera-based attacks. (b) An exploration of the design space of CipherCard concept through the implementation of three prototypes, and (c) a feasibility evaluation and a workload evaluation that validate the usability of CipherCard and its form factors.

## 2 Related Work

Researchers have developed several authentication schemes and interaction techniques with the goal of hindering shoulder-surfing attacks. However, many are still susceptible to camera-based attacks, where the input can be reviewed in detail at a later time.

*Cognitive trap* techniques increase the complexity of the authentication process, for example by showing distractive keystrokes or cursor movements to observers. This makes it more difficult for an observer to derive the password [8, 9] from observing a single authentication instance. However, these schemes are still susceptible to camera-based attacks that afford analyzing multiple authentication instances in detail. In contrast, some schemes allow the user to create a customized 3D gesture (e.g. [4, 10]). These have been demonstrated to be challenging to forge from simple video review; however they are still susceptible to automated video capture and analysis from depth cameras. CipherCard occludes the real password entered into the system, thus resisting camera-based attacks.

*Hiding PIN input* techniques shield against the visibility of input actions. In its simplest form, people can position their bodies/hands to shield their actions from an observer's view [11]; however, most users do not do this [12]. While easy to perform, shielding is vulnerable to well-placed cameras [12] or thermal imaging after entry [13]. Many technical approaches also try to decrease the visibility of password entry. Examples include back-of-the-device PIN entry [3, 6], eye gaze input [14], pressure [11], and haptic/tactile feedback as secret input or output channels to assist password entry [1, 15, 16]. Although resistant to direct human observation, all remain susceptible to video- or audio-based observation attacks. For example, a pressed fingertip can be detected from the change of its color [11]. Similarly, haptic/tactile feedback can be detected from a recorded sound track [16]. CipherCard allows PIN input to be completely hidden from observers, video and audio recording.

*Biometric methods* distinguish users based on biometric characteristics, e.g. fingerprint, hand geometry, retina, etc. [17] or their behavioral signatures [18]. Biometric methods are effective against video-based attacks but suffer major drawbacks preventing them from wide deployment in real-world applications. For example, physiological biometric characteristics are not renewable after the attacker has successfully forged them. Furthermore, behavioral biometric signatures are prone to high recognition error, making them impractical [18].

*Physical token methods* require that the user present a physical object to a reader (e.g. keyfob), and these are immune to shoulder-surfing [19]. However, these require

entry systems to have special hardware to detect the physical object, such as an RFID card readers. Recent advances have begun to explore general capacitive touchscreens as sensors; however this is still in its infancy [20]. Of course, a lost or stolen token could still allow individuals to pass the authentication. This is not the case for CipherCard, which requires both the token and password to pass authentication.

*Two-factor authentication (TFA)* requires a user to present at least two of three factors: knowledge (“something you know”, e.g. password or PIN), inference (“something you are”, e.g. biometric characteristics), and possession (“something you have”, e.g. physical token) [21] to enhance authentication security. Combining a password with a physical token requires a dedicated hardware (e.g. RFID readers) that is usually unavailable on a typical touchscreen device. A popular solution involves asking a user to enter 2 codes (e.g. a PIN + a 1-time passcode generated by a token) through a 2-step authentication process [21]. A common problem in these solutions is that they are vulnerable to man-in-the-middle attacks. Whereas, CipherCard is resistant to such attacks. Additionally, CipherCard does not expose system passwords to attackers. This allows a number of novel applications to be carried out by CipherCard users. For example, a user can use an easy-to-memorize user password while the system can be protected by a strong system password. The user can also use a single user password to authenticate multiple systems each protected by a different system password. Finally, the user can renew a system password while continuing to use the existing user password (see details in Sect. 7).

### 3 CipherCard Concept and Design Space

CipherCard is an opaque overlay that is placed on top of a software authentication UI, acting as a physical proxy to the UI elements. Users can place it on a touchscreen, and use it as a PIN pad to enter PINs. When touching the front side of the card (e.g., a button), the card generates a touch point on its back, however at a different input location on the underlying screen. This creates a randomly preset or user specified permutation between the two sets of locations on the two sides of the card. Thus, the CipherCard provides substitution cipher capabilities, translating the touch input sequence on the CipherCard (*user password*) to another unique sequence that is sent the touchscreen (*system password*). The system password is never exposed. So long as the touchscreen UI and card layout are compatible, which can be achieved through software modification on touchscreen devices, one CipherCard can be reused for an arbitrary number of different PINs for different applications.

The concept behind CipherCard can be realized in a number of different ways. We articulate the factors that describe this design space, and the trade-offs these present.

#### 3.1 Passive vs. Active

CipherCards can be made either passive or active. A *passive* CipherCard translates touch via electrical wiring, and requires no battery or external power source. The passive CipherCard can be cheap to design, produce, and customize for various touchscreen

devices and authentication UIs. It can be disposed and replaced when a user needs a different pattern (e.g. to renew a system password) at a minimal cost. It can also be made reconfigurable (e.g., via jumpers) with more engineering effort and monetary cost.

In contrast, an active CipherCard receives user touches, and uses a control circuit and electrodes to remap those to the touches matching those required by the capacitive sensor on the authentication device. This mapping can be reconfigured through software, giving the user control over the substitution cipher. Furthermore, it is possible to generate complex mappings, where a single touch on the front side generates multiple fake “touches” on the back side (i.e. a *1-to-m* mapping). An active CipherCard has chips, circuits, and software, and thus is more costly.

### 3.2 Input/Output Resolution

The input and output resolution of CipherCard is determined by the number of electrodes on either side of the card. The output resolution of CipherCard is also determined by authentication UIs and the input required. We restricted our early explorations to simple PIN pads of 10 electrodes (to enable 0–9 number entry); however, it is possible to scale CipherCard keypads with more keys, and even to gestural entry, given higher resolution and electrode density.

### 3.3 Form Factor

CipherCards should be easy to carry and deploy. For example, it can resemble a credit card carried in a wallet or purse, or ID tag worn on clothing. Alternatively, it can be integrated into flat daily personal belongings, e.g. a wallet or phone case (that flips open) or even an existing bank or credit card—this avoids the need to carry an extra card. Integrating a passive card into personal belongings can be relatively easy due to its simplicity, but can be challenging for an active card without significantly impacting the normal usage of the personal item. Finally, an active CipherCard can be integrated into existing personal electronic devices, e.g. smartphones or tablets.

### 3.4 UI Alignment

A practical concern is the variety of sizes and layouts of PIN pads may not (by default) match that of a CipherCard. Having a mechanism that can automatically align the two interfaces may largely improve the practicality of the concept.

*Fixed Alignment to PIN pad.* A passive card must be made to match a particular touchscreen layout and is not scalable to UIs with different layouts or button sizes. An active card could utilize higher output resolution to be able to scale to different UI layouts and button sizes (within the card’s physical dimensions).

*PIN pad aligns to CipherCard.* An alternative approach could be to have the software PIN pad align with a CipherCard (either passive or active) automatically. This could be achieved by, for example, adding spatial tags to CipherCard [22]. This would allow the touchscreen to identify the size, position, and orientation of a CipherCard, and align the PIN pad interface accordingly. We expect that the size and

layout of the buttons inside PIN pad boundaries can follow a common standard, thus once boundaries are detected by the touchscreen device, the buttons can be automatically aligned with the electrodes of the CipherCard. In cases when custom sizes and layouts are needed, this information can be pre-stored in the touchscreen devices, and can be loaded upon receiving an encoded ‘request’ from the CipherCard. A request can be encoded into a certain touch patterns (either static or dynamic) based on the pattern of electrodes interpreted by the touchscreen [22].

Finally, the system software can be used to consider only the relative locations (rather than the absolute locations) of the generated touches, where it can match it against a given pattern. That is, the numeric password is treated as a gestural password. While this means that the CipherCard can be placed anywhere on the system screen, security is somewhat reduced as some key combinations may create the same gestural pattern.

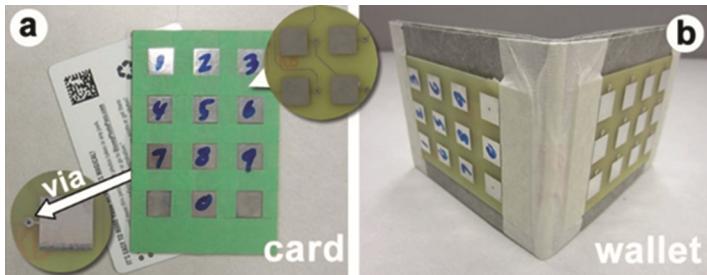
## 4 CipherCard Prototypes

We developed three proof-of-concept prototypes based on this design space: a passive credit card sized prototype; a passive wallet-based prototype, and an active smartphone-based prototype. These prototypes are described below and are the subject of our feasibility and usability studies.

### 4.1 Card-Shaped CipherCard

CipherCard works based on the fact that touch input can be simulated by any conductive object in contact with the capacitive sensor and electrically connected to the user’s hand (or body). We implemented a card-shaped passive CipherCard using a printed circuit board (PCB). Each side of the card contains an identical  $3 \times 4$  grid layout of electrodes (Fig. 2a). For the sake of simplicity, this prototype does not employ utility electrodes for detecting the size and orientation of the card. Each electrode on the front side is uniquely connected to an electrode on the back side. Connections can be either randomized or pre-specified by the user (so they can use a particular user/system password mapping) at the time of manufacture. The electrodes ( $1 \times 1$  cm) and connecting paths (0.025 cm width) were printed using a thin layer of tin. To prevent attackers from deciphering the mapping by visual inspection, CipherCard must be constructed in a manner that hides the connecting paths, e.g. by a surface material or by using a multi-layer PCB design. In our prototype, the connecting paths were covered by paper tape.

To connect the electrodes on both sides, we used tin-coated holes (“vias”). For each electrode on the bottom, we connected it to a via (diameter: 0.2 cm and hole size: 0.071 cm) placed 0.1 cm away from its edge (Fig. 2a). Connecting an electrode to a via from the top connects it to the corresponding electrode on the bottom. Finally, the connecting paths were covered by tape to shield the connection pattern from outside the card. The finished prototype measures  $8.6 \times 5.4 \times 0.15$  cm (L  $\times$  W  $\times$  H), only slightly thicker than a standard credit card, and it can be easily carried in a wallet.



**Fig. 2.** Left: Card-shaped CipherCard with a  $3 \times 4$  electrode grid. Top callout shows the internal wiring and bottom callout shows the via and electrodes on the back. Right: Wallet-based CipherCard.

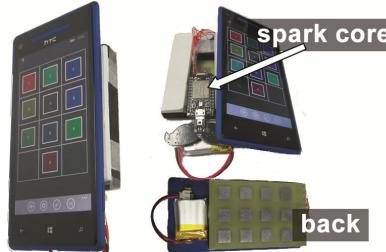
## 4.2 Wallet-Based CipherCard

To demonstrate that CipherCard can be integrated into a daily personal belonging, we built a second passive prototype based on a conventional wallet (Fig. 2b) of size  $10 \times 8.2$  cm when folded. Similar to the card, each side of the wallet has a  $3 \times 4$  grid of electrodes, where one side's electrodes are connected to the opposite side through copper wires. Our prototype uses two hard PCB, but we expect a deployable version to use flexible materials, for example, tin paths printed on PET film. What is important is that the deployable version preserves the appearance, feel and functionality of a wallet.

## 4.3 Smartphone-Based CipherCard

We explored the feasibility of an active CipherCard by creating a smartphone prototype (Fig. 3). We used an HTC 8X Windows Phone as our platform. Input is handled by the phone's native touch input API, and passed via WiFi to a Spark Core development board. The Spark Core drives a  $3 \times 4$  grid of electrodes printed on a plastic boards. A touch is simulated by programmatically connecting one of the pins of the Spark Core to the ground (e.g. configuring the pin as output and set its voltage to 0 V). We found that the ground of Spark Core could not reliably trigger a touch, and thus solved this by connecting the battery jack to the phone body: when the phone is held by user's hand, the Spark Core is grounded through the user's body, and generates simulated touch points reliably. While our prototype is unwieldy, we expect that the deployable version would integrate the logic into the phone hardware, and integrate the simulated touch circuitry into the phone's body.

To simplify our explorations, we constrained the output resolution of our CipherCard prototypes to a fixed PIN layout. However, further engineering efforts would allow resolution of the electrodes to be significantly increased (e.g.  $20 \times 20$   $2 \times 3$  mm electrodes), thus taking advantage of the higher input resolution available from the smartphone.



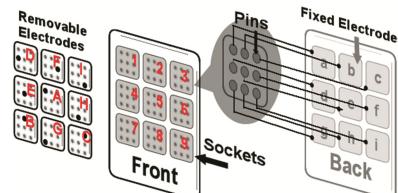
**Fig. 3.** Phone-based CipherCard.

## 5 Reconfigurable CipherCard Mappings

In this section, we present reconfigurable designs for both passive and active CipherCard.

### 5.1 Passive CipherCard Reconfigurable Design

Completely passive cards are cheap to produce. However, they must be replaced when a user needs a different pattern (e.g., to change the desired user or system password). It would be more convenient to design CipherCard so they can be reconfigured on the fly. We designed (not implemented yet) one possibility, illustrated in Fig. 4, which allows a user to reconfigure the connection pattern by rearranging the positions of the electrodes on one side (here the front side). Using a  $3 \times 3$  grid layout as an example, each electrode (numbered *A-I* in Fig. 4) on the front side can be freely removed and re-plugged into any of the 9 sockets (numbered 1–9), and the permutation order they are plugged in intuitively defines how each socket location maps to one of 9 fixed-position electrodes on the back side (numbered *a-i*). To make this possible, we must provide a mechanism to ensure the same removable front-side electrode (e.g., *A*) always connects to the same back-side electrode with the matching letter (*a*), regardless of which socket it (*A*) is plugged into. This is enabled by having 9 small conductive pins ( $3 \times 3$ ) inside every socket. Each pin is hardwired to one of the back-side electrodes with the corresponding relative position (e.g., top-left for *a*). Each front-side electrode has only one pin on its bottom, the relative position of which corresponds to that of the back-side electrode with the matching letter (again top-left for *a*). Thus, whichever socket that electrode *A* is plugged into, its pin always contacts the socket pin that connects to electrode *a*, and so on. Therefore, changing the electrode for a certain socket position will change the position of its associated

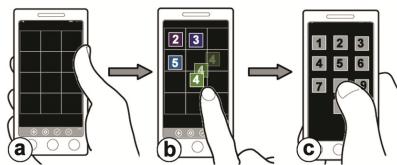


**Fig. 4.** Design of a reconfigurable CipherCard.

touch point seen by a touchscreen. With this simple design, a CipherCard pattern can be reconfigured as easily as switching positions of removable electrodes.

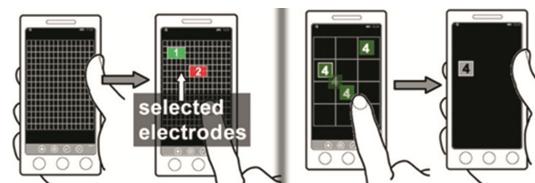
## 5.2 Mobile App for Reconfiguring Active CipherCard

We also implemented a Windows Phone app, which illustrates one interface for reconfiguring the previously described active CipherCard. By default, our app shows a 10 key PIN pad when it starts (Fig. 5a). The touch locations (e.g. electrodes) are shown to guide the reconfiguration of the key mappings. To change a key mapping, the user drags a number key to inside a desired touch location (Fig. 5b). This way, when the key is tapped, the corresponding touch location is triggered (Fig. 5c). Once the configuration is confirmed, the activated touch location is highlighted. When needed, the user can add or remove a number key. To configure a *1-to-m* mapping, the user can duplicate a number key, and then drags each of the duplicated keys to a desired location (Fig. 6 right). Once done, tapping the number key triggers the associated locations in a sequence that the keys were created.



**Fig. 5.** (a) grid shows the position of the electrodes; (b) dragging a number key to inside a desired electrode to change a key mapping; (c) finished configuration.

Although not implemented in the hardware, our software also supports configurable key sizes and layouts. First, a variety of standard keyboard layouts can be selected, where each matches the keys and layout of a particular touch-screen security system. Second, layouts can be designed from scratch, although the interactions to do so are more complex. For example, the user can specify the dimensions of the key, and drag it to a desired location. The software also allows the user to scale key sizes using pinch gestures. Our software automatically identifies the candidate location(s) that need be triggered for simulating a touch at the position of the key. To do so, the user first specifies the resolution of the touch locations (or electrodes). The software then walks through the locations and associates one with the key, which has the largest overlap with that key (Fig. 6 left). Notice that most of the capacitive sensors ignore touches that are smaller



**Fig. 6.** Left: configuring a grid of  $15 \times 15$  electrodes; Right: duplicating a number key to configure a *1-to-m* mapping.

than a threshold size (e.g. 3 mm for Microsoft Surface). To accommodate this, our algorithm triggers all the electrodes (if smaller than 2 mm) that reside inside the key.

### 5.3 Authoring a New Card

At the current stage, passive CipherCards can be designed using popular circuit software (e.g. Altium), and built using a standard PCB. A much easier way is to print on paper using a home printer and conductive ink [23]. This allows CipherCard to be widely adopted for home and office use.

## 6 Security Analysis

Our assumption of a threat model is based on real world threats, under which a shoulder-surfing attack may take place. We assume the user is in a public environment fully controlled by the attacker, who has hidden a number of high-resolution cameras in that environment. The cameras can record (from multiple angles) all the user's actions on a software PIN pad on a capacitive touch-sensing device. Multiple authentication sessions of the same user can be recorded, which can then be reviewed in order to extract the PIN. We also assume that the adversary has direct access to the authentication system but without the possession of CipherCard.

An attacker who has observed the user password is unable to pass authentication without possessing the user's CipherCard. Copying the card configuration is also difficult without physically possessing it. CipherCard can be deployed where the system password is not revealed to the user. This means the user would only know their user password, which in turn mandates CipherCard in user verification. This makes the user immune to the social engineering attacks [24], where the user may divulge the system password to attackers, who can then bypass the CipherCard to authenticate directly.

Losing both the user password and a CipherCard to an adversary will grant access to protected services or locations. Hacking into an active card, e.g. user's phone, or breaking into a computer which stores translation files may disclose a user's password mapping to an adversary. Individuals or organizations that design or have access to the design of the password mapping may also present risks to the security of CipherCard.

CipherCard does not prevent attacks directly on the authentication device. With the *1-to-1* translation of a password, CipherCard does not increase the overall entropy (i.e. the total number of possible authentication inputs seen by the system) [25]. We therefore assume the original password mechanism on the device is sufficiently strong on its own (e.g., with an appropriate password length and a limited number of trials) against direct attacks, e.g. brute-force attacks (i.e. enumerating all possible passwords) [26]. CipherCard however protects against dictionary attacks – a user may choose a common word as the user password, yet the translated system password is highly unlikely to be in the dictionary of guessed passwords. The entropy of the *1-to-1* mapping is equal to the total number of permutations of the  $n$  electrodes, i.e.  $n!$ . For example, a  $3 \times 3$  grid layout offers  $9! = 362880$  unique CipherCard patterns. In contrast, a *1-to-m* translation of a password increases the overall entropy by allowing a longer system password, thus providing a higher level of security. Note that  $m$  can vary for each character in the user

password. Brute-force attacks can be extremely difficult if the length of the system password is unknown to the attacker, and can be thwarted simply by limiting the number of incorrect entries and/or by introducing time delays between attempts. Notice that if a chosen user password is too easy to guess, the level of security of CipherCard can be reduced. For example, in an extreme case of *1-to-m* mapping, a user password can contain only one character, which serves as a shortcut to a longer system password. This way, the user password becomes extremely easy to obtain. Even so, the attacker would still need to somehow take possession of the CipherCard.

## 7 CipherCard Usage

CipherCard makes it possible for people to choose easy-to-remember PINs instead of using strong ones that are less memorable, which often imposes security concerns [3]. Furthermore, CipherCard makes it possible to have a single user password associated with different system passwords, either through the use of multiple passive CipherCards set to that user password (but generating different system passwords), or a single active CipherCard (which would change the mapping based on the service). This allows the user to reuse passwords for multiple accounts without significantly impacting security [27].

*Changing System Passwords.* Renewal of passwords is commonly enforced by organizations to enhance security, but places undue burden on users to generate or remember new passwords. CipherCard allows for refreshed system passwords while allowing users to continue using their existing user passwords in two ways: they can use a new (passive) CipherCard, or the internal mapping in the CipherCard can be changed. Either way, user password now maps to a new system password without loss of security.

*Changing a User Password.* Many systems allow the user to change their system password after they have correctly logged in. After selecting that option, a user can simply place the CipherCard on the authentication UI, and enter a new user password which, in turn, generates the new system password.

*Setting a User Password Based on an Existing System Password.* In many scenarios, a system password is shared by many people (e.g. door entry); in this case, it is desirable to keep the system password, but also allow for a mapping between this and a user-chosen password. Because different CipherCards can generate the same system password from different user passwords, this becomes easy to do with our reconfigurable designs. The actual mapping between the two can be done by any of the previously described methods.

*Replacing a Lost or Damaged CipherCard.* A new CipherCard can be authored and granted to the user if the original one is lost or damaged. The user can also make one at home (see 5.3). The design of the password mapping needs to be securely stored on a safe computer in order to preserve the security of CipherCard.

## 8 Feasibility Study

We conducted two studies to evaluate the concept of CipherCard, our designs, and identify potential usability issues. In our first study, we used heuristic evaluation to identify usability issues of the three prototypes. We were also interested in perceptions of the security of the scheme. We did not implement UI alignment for CipherCard. This allowed us to investigate the feasibility of actual deployment without modifying existing touchscreen devices. While we believe that CipherCard warrants a long-term field deployment study, in this early development stage, we deem this study a necessary step towards refining and improving the concept before they can be deployed and studied.

### 8.1 Participants

Our heuristic evaluation was conducted with usability experts. We recruited five professional usability engineers (25–40 years old) from industry. Two participants had one year of industry UX experience, one had >3 years, and two had >5 years of experience.

### 8.2 Apparatus and Procedure

At the beginning of the study, we showed the participants the three CipherCard variations, e.g. card-shaped, wallet-based, and phone-based CipherCard. We then walked them through three CipherCard usage scenarios: entering a PIN into (1) a touchscreen door lock, (2) a public kiosk, e.g. ATM and POS terminal, and (3) a personal mobile device (e.g. a tablet). To simulate the ATMs or door locks, we used Microsoft Surface tablets positioned in different ways. For example, to simulate a door lock, the tablet was hung on a vertical surface. To simulate a public kiosk, the tablet was tilted 35° on desk. To simulate a mobile scenario, the participant was asked to hold the tablet using their non-dominant hand and authenticate using CipherCard with the other hand. For each usage scenario, the participants were asked to enter a 4-digit PIN into a PIN pad application running on the tablet. After a PIN was entered, the application indicated whether the authentication succeeded or failed. In order to

Participants were encouraged to put themselves in the mindset of someone using these systems in real-life usage situations, e.g. taking the card from their pocket before use. They were allowed to try and use the prototypes for as long as they wanted prior to completing a questionnaire (7-point Likert scale) and an interview, in which the participants were asked about their perceived security and portability of CipherCard as well as their mental demand, physical demand, effort, frustration, and concentration when using the prototypes.

### 8.3 Results

Overall, the participants welcomed CipherCard as a method to resist camera-based shoulder-surfing. Their feedback confirmed the merits of the prototypes, e.g. security and portability, but also identified issues that may cause cognitive overhead (Fig. 8 Left).

The results reported below are using median.

**Merits of CipherCard.** *Security.* All of the participants perceived CipherCard as more secure against shoulder-surfing than current practices (6, 7 being the most secure; s.e. = 0.5), e.g. directly entering the PIN. For the participants, who had expressed interest of using CipherCard (e.g. P1&P5), they found it highly attractive to have an extra layer of security. Some of the positive comments included “*I shield my PIN entry, it is my habit but I don’t feel I have to (with CipherCard)*” -P1 and “*I see myself using CipherCard to unlock my door because now I have the security of a bankcard, if someone wants to break into my house, they need to get my card as well.*”-P5.

*Portability.* All prototypes were rated highly portable, e.g. card: 7, wallet: 7; phone: 7, with 7 being strongly agree; all s.e. = 0) regarding the convenience in carrying them around. P1 commented that it would be convenient to carry the phone-based CipherCard because “*it is something I carry around anyways.*” The wallet received similar comments, e.g. “*I don’t have to carry something else as I already carry one*”-P5. While the card-shaped CipherCard is considered an extra burden (i.e. a new thing to carry), our participants found it easy to carry as well: “*I have a lot of cards anyways, so I don’t think if carrying a lot of them (cards) will be an issue*” -P1, and “*I will be ok to carry it around if the credit card company decides everybody needs to do*”-P3.

**Issues that Cause Cognitive Overhead.** *UI alignment.* Prior to entering the PIN, CipherCard needs to be physically aligned with authentication UI, e.g. PIN pad. This was seen as an unwanted extra step. Among the three prototypes, the card-shaped design was the easiest to align, while the rest were initially challenging for the first time users. Misalignment resulted in touches being unregistered on the touchscreen, which had consequently caused frustration.

*Slippery Screens.* Touchscreens are slippery. This had made alignment even more difficult. The participants had to spend extra effort when holding the prototypes steadily, especially when the screen was tilted. The participants also worried about dropping their phone on tilted screens.

*Two-Handed Operation.* Entering a PIN on the prototypes while making sure the device did not slide required using two hands. Two-handed operation introduced unnecessary effort for the participants to prepare for using the card. For example a participant commented that, by requiring two hands, “*I will have to put my bag down and use both hands to operate*”-P1. Additionally, the holding hand had sometimes occluded the number buttons that the participant wanted to tap.

*Orientation.* The translated output locations are dependent on the orientation of the card and the side that is used for input. The phone- and card-based prototypes have clear visual affordance, making it easier for the participants to identify the desired side and orientation to use. However, the wallet is symmetrical in its appearance, thus requiring extra effort from the participant to figure out the right direction.

*Preparation Effort.* All of the aforementioned issues had introduced unnecessary preparation efforts from the users prior to entering the CipherCard’s user password. Overhead

also includes the effort to take out the device from where it is carried. The card-shaped device is less convenient than the other two prototypes in the sense that the users will have to take out the wallet first (assuming the card is not worn as an id tag). With the phone, the participants even more overhead, where they had to first unlock the phone, open the app, and then search for the desired card mapping.

## 9 Improved Design

After carefully reviewing the results from the first study, we came up with a number of solutions to resolve some of the most outstanding issues.

To reduce the preparation time for phone-based CipherCard, we implemented a new function, which allows the phone to automatically load a desired mapping by tapping it on a Near Field Communication (NFC) tag. In circumstances that the size and layout of the capacitive PIN pad does not match the one on the phone, the software can automatically load a key pad configuration that matches its specification.

For issues regarding UI alignment, slippery screens, and two-handed operation, we designed a card-holder that can be attached on top of the capacitive PIN pad. This allows the user to snap CipherCard into the right position on the screen without aligning or holding by hand. The card-holder guides the position of CipherCard, holds it onto the screen, and aligns it properly. We implemented a prototype on-screen card-holder to demonstrate the idea (Fig. 7). To use it, the user simply slides the card-shaped prototype into the holder from the top and enters a PIN. This allows single-handed operation without the need for user alignment. Alternatively, magnets can be attached to the card and screen to achieve the same goal, while preserving the flatness of the screen. This design is more suitable to phones and wallets as they do not have a uniform form factor. Thus, an on-screen holder may not work for them. The user can snap CipherCard onto the screen (e.g. using magnets) in an arbitrary orientation. The software PIN pad aligns with CipherCard automatically. It can even adjust its button size and layout to fit those of CipherCard (see details in Sect. 4). We can envision many different approaches for developing the snap-in mechanism, exploring which are outside the scope of this paper. We thus leave them for future work. Instead, we focused our investigation on the effectiveness of our improved designs.



**Fig. 7.** Snap-in card holder allows one handed PIN entry.

## 10 Preliminary Workload Evaluation

The goal of this study was to verify the concept of our improved design as well as the overall usability and security of CipherCard. To focus on the concept rather than the implementation, we mocked up the snap-in mechanism using the on-screen card holder shown in Fig. 7. For the wallet and phone, we used double-sided tape (on the back of different CipherCards) to simulate a magnetic snap-in effect.

### 10.1 Participants

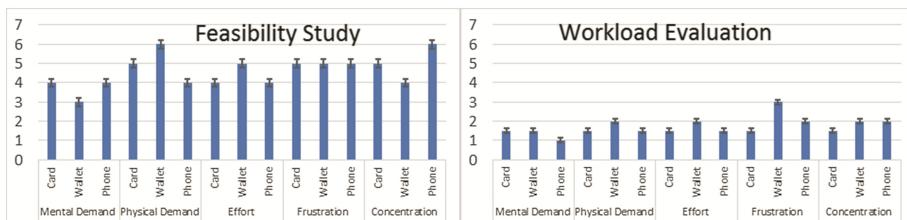
Six participants (5 males and 1 female). All were adult office workers with prior experience using PIN pads and TFA.

### 10.2 Apparatus and Procedure

The procedure is similar to the feasibility study, except with the phone-based prototype, the participants were asked to tap the phone on a NFC tag to load the app before entering a PIN. Participants were trained on the use of the snap-in guide for the card prototype. For the wallet- and phone-based prototypes, we asked participants to snap them onto the software PIN pad and imagine that alignment would be automatically adjusted.

### 10.3 Results

*Reduced Cognitive Overhead.* Overall, the participants rated cognitive workload being very low (1.5, with 1 being extremely low; s.e. = 0.1). The snap-in solution led to low mental/physical demand, effort, frustration, and concentration (Fig. 8 Right). The result of this study indicated the importance of having the snap-in feature before CipherCard can be deployed.



**Fig. 8.** Average responses on several measures of cognitive overhead from both studies (Likert-scales: 1: very low, 7: very high) (Error bar shows standard error).

Participants found the wallet and phone had higher level of frustration than the card due to their cost value and the potential danger of exposing them in the public. For example, “*In places that is not safe, I don’t want to pull out my wallet because muggings are really common, and there is way too much personal information in the wallet*”-p7. Although, tapping a NFC tag still requires extra effort from the users, participants found it much easier to do than searching through an application list.

## 11 Results from Both Studies

*Security.* The results from both studies confirmed that CipherCard was perceived more secure against shoulder-surfing than conventional PIN entry. When asked if they felt comfortable not knowing their system password, a slight majority of participants (6/11) said they preferred knowing it. Although all understood the security benefit of not knowing the system password (e.g. enforce TFA or against social engineering attacks), more participants preferred to have a backup in case a CipherCard was lost, stolen or otherwise not available. Seven of the 11 participants expressed interest in using CipherCard in a public kiosk or door lock. A minority (4) expressed interest in using CipherCard on mobile devices for highly secured application, e.g. online banking. Others were less interested, feeling they had more control hiding their input on a mobile device.

*Maintaining Multiple Passwords.* Overall, the participants saw the merits of using CipherCard to help release the workload of memorizing multiple strong passwords, e.g. 6 with 7 being strongly agree. When choosing between using one CipherCard with multiple user passwords and multiple CipherCards with a single user password, all participants leaned towards using one card. They explained it would be easier than carrying multiple cards. More participants (7) also leaned towards getting a new (passive) card when renewing a system password, as they could benefit by keeping the current user password (assuming the card is associated with only one PIN).

*Social Pressure.* Participants were asked to rate how much social pressure they may feel when using CipherCard in front of stranger, friend, and family, where the viewer may perceive it as an insult. They did not feel social pressure using CipherCard in front of strangers, friend, and family (all 7, with 7 being strongly disagree that they felt social pressure; s.e. = 0); They did not think they would feel uncomfortable if others (e.g. stranger, friend, or family) were to use CipherCard in front of them (1, with 7 being strongly mind; s.e. = 0).

## 12 Discussion and Limitations

In the section, we discuss the insights and limitations we discovered from our own experiences designing CipherCard.

*Change of Authentication Behavior.* While CipherCard does not change the way a user enters a PIN, it changes a user's authentication behavior, i.e., it requires the user to carry a card and put it on the touchscreen prior to entering a PIN. Users may be resistant to this extra work. Like any other security system, users are always the key to ensure the success of CipherCard. While people have been found to be the 'weakest link' in the computer system [24], their security behavior can be changed through education and proper design of security systems. We see that the features such as allowing easy-to-memorize user password, reuse of a user password for multiple authentication systems and for renewing a system password are handy trade-offs that may encourage users' authentication behavior by actively using CipherCard.

*Convenience vs Security.* Our study showed that people do understand the importance of security. However, it is often the case that people sacrifice security for the sake of control or convenience [28]. CipherCard tries to motivate user’s security behavior (e.g. using TFA) by providing a set of handy features. While welcomed by our participants, users need to be aware that some of the features may introduce potential security risks. For example, using a single user password for multiple accounts or updating CipherCards but never changing user password may reduce the security of CipherCard. In addition, if an adversary steals a wallet containing multiple CipherCards, all with the same user password, he will be able to access *all* associated accounts if he knows that password, even though their system passwords may differ. Future work needs to focus on convenient techniques without impacting security.

*Modification of the Existing Authentication Device.* The UI alignment technique needs to be developed before CipherCard can be deployed in the field. This, however, requires minor augmentation of the existing hardware and/or software, which would increase the cost of deployment.

*Applications.* We demonstrate CipherCard on capacitive touchscreen PIN pads but we envision the concept can be applied to popular gestural and QWERTY keyboards.

*Study.* CipherCard warrants a long-term field study, which will help in understanding its practical usability in real-world use. The results from a field study might be more nuanced from the results from a laboratory environment due to artificial setups [12].

*Prototypes.* Our prototypes were designed as proof of concepts. Deployable systems will need more attention to how CipherCards appear, the cost of manufacture, and the reliability of the electronics.

## 13 Conclusion

In this paper, we introduced the concept of CipherCard to prevent PIN entry on capacitive touchscreens from camera-based shoulder-surfing attacks. CipherCard remaps a user’s touch point to a different location on the touchscreen, thus translating the visible user password into a hidden system password received by the touchscreen. We explore the design space of CipherCard, and implemented three proof-of-concept prototypes. We evaluated the CipherCard concept with two user studies. The first study identified several usability issues, where we then proposed solutions that were the subject of the second study. User feedback from both studies confirmed the promise of CipherCard. Those studies (and our own experiences) also revealed various issues and tradeoffs that could affect its acceptance, its real-world use, and that should be considered in evolving designs. Of course, we are still in the early stages. Future work will evolve CipherCard’s design, ideally resulting in a field deployment form which real-world usage data and its practicality can be better understood.

## References

1. Bianchi, A., Oakley, I., Kostakos, V., Kwon, D.S.: The phone lock: audio and haptic shoulder-surfing resistant PIN entry methods for mobile devices. In: TEI 2011, pp. 197–20 (2011)
2. Kumar, M., Garfinkel, T., Boneh, D., Winograd, T.: Reducing shoulder-surfing by using gaze-based password entry. In: SOUPS 2007, pp. 13–19 (2007)
3. Luca, A.D., Harbach, M., Zezschwitz, E.v., Maurer, M.-E., Slawik, B.E., Hussmann, H., Smith, M.: Now you see me, now you don't: protecting smartphone authentication from shoulder surfers. In: CHI 2014, pp. 2937–2946 (2014)
4. Shirazi, A.S., Moghadam, P., Ketabdar, H., Schmidt, A.: Assessing the vulnerability of magnetic gestural authentication to video-based shoulder surfing attacks. In: CHI 2012, pp. 2045–2048 (2012)
5. Wiedenbeck, S., Waters, J., Sobrado, L., Birget, J.-C.: Design and evaluation of a shoulder-surfing resistant graphical password scheme. In: AVI 2006, pp. 177–184 (2006)
6. Luca, A.D., Zezschwitz, E.v., Nguyen, N.D.H., Maurer, M.-E., Rubegni, E., Scipioni, M.P., Langheinrich, M.: Back-of-device authentication on smartphones. In: CHI 2013, pp. 2389–2398 (2013)
7. Vogel, D., Baudisch, P.: Shift: a technique for operating pen-based interfaces using touch. In: CHI 2007, pp. 657–666 (2007)
8. Kim, S.-H., Kim, J.-W., Kim, S.-Y., Cho, H.-G.: A new shoulder-surfing resistant password for mobile environments. In: ICUIMC 2011, no. 27 (2011)
9. Tan, D.S., Keyani, P., Czerwinski, M.: Spy-resistant keyboard: more secure password entry on public touch screen displays. In: OZCHI 2005, pp. 1–10 (2005)
10. Kratz, S., Aumi, M.T.I.: AirAuth: a biometric authentication system using in-air hand gestures. In: CHI 2014 EA, pp. 499–502 (2014)
11. Kim, D., Dunphy, P., Briggs, P., Hook, J., Nicholson, J.W., Nicholson, J., Olivier, P.: Multi-touch authentication on tabletops. In: CHI 2010, pp. 1093–1102 (2010)
12. Luca, A.D., Langheinrich, M., Hussmann, H.: Towards understanding ATM security: a field study of real world ATM use. In: SOUPS 2010, pp. 1–10 (2010)
13. Mowery, K., Meiklejohn, S., Savage, S.: Heat of the moment: characterizing the efficacy of thermal camera-based attacks. In: WOOT 2011, p. 6 (2011)
14. Luca, A.D., Denzel, M., Hussmann, H.: Look into my eyes!: can you guess my password? In: SOUPS 2009, no. 7 (2009)
15. Sasamoto, H., Christin, N., Hayashi, E.: Undercover: authentication usable in front of prying eyes. In: CHI 2008, pp. 183–192 (2008)
16. Luca, A.D., Zezschwitz, E.v., Hußmann, H.: Vibrapass: secure authentication based on shared lies. In: CHI 2009, pp. 913–916 (2009)
17. Liu, S., Silverman, M.: A practical guide to biometric security technology. *IT Prof.* **3**, 27–32 (2001)
18. Sae-Bae, N., Ahmed, K., Isbister, K., Memon, N.: Biometric-rich gestures: a novel approach to authentication on multi-touch devices. In: CHI 2012, pp. 977–986 (2012)
19. Roth, V., Schmidt, P., Güldenring, B.: The IR ring: authenticating users' touches on a multi-touch display. In: UIST 2010, pp. 259–262 (2010)
20. Vu, T., Baid, A., Gao, S., Gruteser, M., Howard, R., Lindqvist, J., Spasojevic, P., Walling, J.: Distinguishing Users with Capacitive Touch Communication. In: Mobicom 2012, pp. 197–208 (2012)
21. Schneier, B.: Two-factor authentication: too little, too late. *Commun. ACM* **48**, 136 (2005)

22. Yu, N.-H., Chan, L.-W., Lau, S.-Y., Tsai1, S.-S., Hsiao, I.-C., Tsai, D.-J., Cheng1, L.-P., Hsiao, F.-I., Chen, M.Y., Huang, P., Hung, Y.-P.: TUIC: enabling tangible interaction on capacitive multi-touch displays. In: CHI 2011, pp. 2995–3004 (2011)
23. Kawahara, Y., Hodges, S., Cook, B.S., Zhang, C., Abowd, G.D.: Instant inkjet circuits: lab-based inkjet printing to support rapid prototyping of UbiComp devices. In: UbiComp 2013, pp. 363–372 (2013)
24. Orgill, G.L., Romney, G.W., Bailey, M.G., Orgill, P.M.: The urgency for effective user privacy-education to counter social engineering attacks on secure computer systems. In: CITC5 2004, pp. 177–181 (2004)
25. Burr, W.E., Dodson, D.F., Polk, W.T.: Electronic Authentication Guideline. NIST, USA (2012)
26. Kim, D., Solomon, M.: Fundamentals of Information Systems Security. Jones & Bartlett Learning, MA (2010)
27. Von Zezschwitz, E., De Luca, A., Hussmann, H.: Survival of the shortest: a retrospective analysis of influencing factors on password composition. In: Kotzé, P., Marsden, G., Lindgaard, G., Wesson, J., Winckler, M. (eds.) INTERACT 2013, Part III. LNCS, vol. 8119, pp. 460–467. Springer, Heidelberg (2013)
28. Cranor, L., Garfinkel, S.: Security and Usability. O'Reilly Media, CA (2005)

# Digital Signage Effectiveness in Retail Stores

Mari Ervasti<sup>1</sup>, Juha Häikiö<sup>1</sup>, Minna Isomursu<sup>2</sup>, Pekka Isomursu<sup>3(✉)</sup>,  
and Tiina Liuska<sup>3</sup>

<sup>1</sup> VTT Technical Research Centre of Finland, Oulu, Finland

{Mari.Ervasti,Juha.Haikio}@vtt.fi

<sup>2</sup> Department of Information Processing Science, University of Oulu,  
Oulu, Finland

Minna.Isomursu@oulu.fi

<sup>3</sup> Department of Media, School of Media and Performing Arts,  
Oulu University of Applied Sciences, Oulu, Finland

Pekka.Isomursu@oamk.fi, Tiina.Liuska@gmail.com

**Abstract.** This paper presents results from a study on the effectiveness of digital signage in the retail environment. The goal of the study was to examine design parameters relevant to digital signage content design which could be used to create guidelines and templates for designing effective digital signage content. In this study, we focused on how video and animation affect the effectiveness of digital signage. When comparing still content with content enhanced with video or animation, no significant difference in effectiveness could be observed. This observation contradicts with earlier studies. Our study supports the views that the digital displays are currently most useful and effective to the younger generation, and that male customers consider digital displays in a store more useful than females do.

**Keywords:** Digital signage · User study · Retail store · Media management · Digital content design · User interfaces · User experience

## 1 Introduction

Digital signs or displays are currently used widely for delivering information in public spaces. Both in- and outdoors, they are ubiquitous in transportation systems, sport stadiums, shopping centers, health care centers and stores. Among other things, digital signage is used for advertising, information sharing and entertainment purposes. The underlying goals of such signage are to provide better service for customers and promote sales. From a retail perspective, digital signage can be seen as one of the many channels that aim to capture people's attention and affect their customer behavior. Digital signage provides retail actors with a channel where content can be updated rapidly and with ease [5]. It can also provide retail actors a communication channel which is entirely in their control, creating opportunities not available with traditional media.

The study presented in this paper examines how effectively customers perceive various visual contents on digital displays in retail stores. We examine if the

effectiveness of digital display content in retail context can be improved with video or animation, and if there is a difference between genders and age groups on the perceived usefulness and effectiveness of digital signage. We also examine how content design for digital signage could be supported by visual design templates that partially automate and unify the content design process. Ultimately, our goal is to understand the interplay between digital signage technology and digital advertisement content, and how they can be combined to achieve the desired impact on retail consumer choices.

The motivation to our study comes from the participating retail company. The company has made a strategic decision to move more of its advertisement content from traditional, external advertisement media, such as newspaper, radio and television, to digital media platforms which are in their direct, real-time control. These platforms include digital signage placed in the company's retail premises, and social media. This change has effects in the operation and management of their advertisement department. The new digital channels provide improved possibilities for real-time content updates and content localization. However, utilizing these possibilities requires new skills, design approaches and content management procedures, as well as thorough understanding of the new platforms.

Our study was executed in two parts. First, we conducted a literature survey on previous digital signage research. Based on our findings, we then executed a field study where we created and delivered digital content to the displays in retail stores of a co-operative retail chain located in Northern Finland and evaluated the effectiveness of different content designs. We used the number of customers who remembered the fact(s) shown on digital displays as a measure for effectiveness.

## 2 Related Work

Previous studies, such as Huang et al. [4] and Höffler and Leutner [3], show that the format of the digital content affects the effectiveness of digital displays. People find video content more attractive and it captures the eye longer than text or still images [4]. In addition, video-based animations seem to be superior to computer-based animations [3], as video content captures the eye somewhat longer [4].

Similarly to banner blindness on web sites, which has been discussed, for example, by Nielsen [9], most public displays are ignored by many people or receive only very few glances [4]. Thus, the process of selective attention also applies to digital signage, an effect known as 'display blindness' [7]. People expect uninteresting content, which leads to a tendency to ignore the displays. In a study conducted by Müller et al. [7], colorfulness was the most important factor influencing whether people thought they would look at the displays.

Shoppers are most responsive to localized information, such as information on new items, promotions and seasonal information, as well as messages about hedonic products, such as food and entertainment [1]. Additionally, shoppers are most interested in messages that address the task at hand and their current need state [1]. According to [10], males are more attracted and receptive to digital signage than females.

### 3 Research Methods

#### 3.1 Study Setup

The pilot study was set up in four different stores in the city of Oulu, Finland. The study duration was two weeks. Three of the stores were local grocery stores and one was a supermarket. Digital content was shown to customers in three of the stores. The fourth grocery store did not have digital displays and we used it as a reference case. Content design was a collaborative effort between retail experts, digital signage experts and researchers. For the study, we created three content themes which were identified by the retail chain to be representative examples of advertisement content they would like to communicate through this media channel: (1) *Energy savings*, (2) *Local food*, and (3) *Familiarity with the store manager* (Table 1). For each theme, we presented the same content using four different types of templates:

- A *Still picture*: nothing was animated (Fig. 1);
- B *Background animation*: there was some animated movement in the content background to draw the customers' attention (Fig. 2a);
- C *Fact animation*: animated movement was added to the key fact in the conveyed message (Fig. 2b);
- D *Narrative animation*: a sequence of animations was used to draw attention to different parts of the message in a desired order.

The abovementioned content themes were presented on the store displays by using one type of content template at a time. In other words, during one pilot study day the themes were shown to a specific store's customers using only A, B, C, or D from the list above. We collected data through customer interviews about the impact that each template had on the store customer. We created the above-mentioned templates partially based on the findings on visual efficiency from related digital signage research but also due to our study goal of experimenting with content formats that could potentially be used as templates to help marketers create new digital content.

Several planning sessions between retail experts, digital signage experts, a media designer and researchers were conducted in order to create an appropriate set of questions for customer interviews in stores. The question set was kept short as customers are often quite busy while grocery shopping. In addition, a public space is not the best possible place for long interviews.

**Table 1.** Study setup in each store.

Store	# of interviews	# of displays	Content themes on displays
Grocery 1	122	2	Energy, local food & store manager
Grocery 2	142	1	Energy, local food & store manager
Grocery 3	43	0	No content (For Reference)
Supermarket	244	8	Energy, local food
<b>Total</b>	<b>551</b>	<b>11</b>	



**Fig. 1.** An example of display content. The template is Still picture: nothing is animated. The content theme is (1) Energy savings. The text says: ‘We are saving energy in this store - up to 50 % - using new energy solutions’.



**Fig. 2 a&b.** Examples of display content. In Fig. 2a the template is Background animation: the branches and leaves grow in the background. The content theme is (2) Local food. The text says ‘Best food is local – Rönkä’s products now in our store – Follow this sign > Local product’. In Fig. 2b the template is Fact animation: the text “Tervetuloa omaan kauppaan!” (i.e. Welcome to your own store!) appears when the store manager turns and points to her upper right. She also briefly looks further down at her name “Sari”. The content theme is (3) Familiarity with the store manager. The text says ‘Welcome to your own store! – Wishes Sari, the store manager’.

All digital displays utilized in the pilot were located indoors. In the grocery stores the displays were typically located quite close to a cash desk. In the supermarket they were located along the corridors. The displays were located quite high with the exception of one small display that was located on a lower level nearby a cash desk that was used only during the peak hours. All displays had rather large screens (about 50 inches), excluding the small display mentioned above.

### 3.2 Study Execution

The customer interviews in stores were conducted during the stores’ opening hours between noon and 5 PM. The length of each interview was from five to ten minutes. The interviewers asked the customers to answer the survey questions when they were leaving the store, i.e. after being exposed to the store’s digital display(s). The total number of interviews was 551. Of the interviewees, 278 (50.5 %) were females and 273 (49.5 %) were males.

Table 1 also presents the content themes used in each store. Questions related to the familiarity of the store manager were not asked in the supermarket since the manager is

not meant to be as visible to the customer in the supermarket as s/he is in the grocery stores. Tablet computers were used by the researchers during the interviews to record and store the answers. Afterwards, the answers were exported to an Excel sheet and analyzed with Excel's analysis tools.

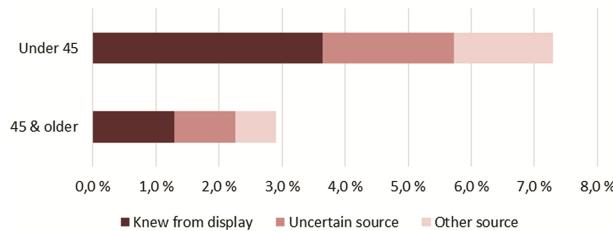
## 4 Key Results and Findings

Our main goal was to collect data through customer interviews about the impact that each content template had on the store customer. Our data shows that none of the templates was very effective, i.e. the majority of the customers did not pay much or any attention to the displays regardless of the template used. This is illustrated by reoccurring customer comments such as "*Where are the digital displays in this store?*" There were 551 respondents in our study. Our results show that there was no statistically significant difference (p-value less than 0.05) on the effectiveness of the displays presenting content with or without video or animation. This is contradictory to the findings of Huang et al. [4] and Höffler and Leutner [3].

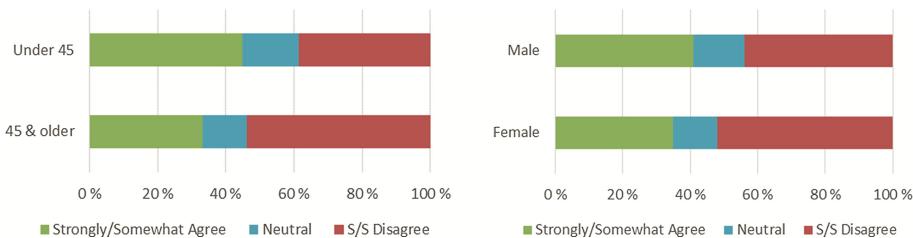
Data from our study supports the view that the digital displays are most useful to or perhaps best exploited by the younger generations. Respondents that were under 45 years old knew the displayed fact on energy saving and recognized the store manager clearly more often than the older respondents (Fig. 3). 4–6 % of the participants that were under 45 years old learned the fact about energy saving from the info displays, whereas only 1–2 % of the participants that were 45 years or older learned this fact.

The respondents' own views were in accordance with our abovementioned observation that digital signage is most useful to the younger generations. To find out about the perceived usefulness of digital displays in the store, the interviewees were asked to evaluate the statement 'I find digital displays in the store useful'. A 5-point Likert scale (1 = Strongly agree, 5 = Strongly disagree) was used to determine the customers' subjective conception. When comparing responses between younger (under 45 years old) and older respondents (over 45 years old) we can see that the younger generation considers digital displays located in a store more beneficial (Fig. 4a). Six responses without age information were excluded from the analysis. The findings discussed above indicate that digital signage, although currently vastly ignored by the customers, may become an increasingly efficient form of media in the future as digital literacy increases generation after generation.

According to our research data, males considered digital displays in a store slightly more useful than females (Fig. 4b). 41 % of male and 35 % of female respondents selected either 'Strongly agree' or 'Somewhat agree' in response to the statement 'Digital displays are useful for me in the store'. 44 % of males and 52 % of females disagreed somewhat or strongly with this statement. Ravnik and Solina's [10] research also supports the observation that males are more attracted and receptive to digital signage than females.



**Fig. 3.** Age distribution of respondents who knew the fact about energy management.



**Fig. 4 a&b.** Figure 4a shows age preference in relation to ‘I find digital displays in the store useful’, and (Fig. 4b) Gender preference in relation to ‘I find digital displays in the store useful’.

## 5 Considerations on Data Validity and Interpretation

Our research method was strongly based on whether the interviewee knew the fact presented through the digital display or not. In the case of a correct answer, we asked where the interviewee thought he or she had learned the fact in question. However, the correctness of their answers is subject to interpretation. As an example, we consider the question about how much energy the store had saved. 14 out of 192 (i.e. 7.3 %) respondents that were under 45 years old knew the correct answer. Half of them said they got the information from the digital display and an additional four were uncertain of where they got the information (‘cannot say’, ‘a guess’ or ‘saw it somewhere’). Three respondents clearly named a different source. Thus, we obtained the result that 4–6 % (i.e. 7–11/192) of the participants under 45 years old learned the fact from the displays in the store. However, it is difficult to evaluate how reliably people can recall in an interview situation from which source they learned certain facts.

A high percentage (39 %) of the respondents identified the store manager from the picture shown to them, and 4.5 % knew the manager’s name. Younger customers knew the manager’s name more often than older customers. However, it is likely that there is distortion in this data due to the fact that many customers already knew or were somewhat familiar with the store manager, which appears to be a quite common situation in smaller grocery stores. Indeed, many of the respondents reported that they identified the store manager from the picture since they had seen the manager in the store. As the energy policy of a particular store was not at all as visible to the customers as was the store manager, we consider energy savings to better indicate how well the customers learned the fact from the display.

## 6 Discussion and Future Work

Our results indicate that many users did not pay attention to digital displays or ignored them altogether. The content format did not have a statistically significant effect on effectiveness, i.e. it was equally low with still, animated and video based templates that we used.

Although the percentages of the correct responses regarding the content of the digital displays may seem small, they are roughly tenfold in comparison to the average click-through-rates (CTR) in display advertisement (i.e. 0.1–0.3 % according to Chaffey [2] and Stern [11]). Also, cost per impression (CPI) in digital displays can be very low if the displays are owned and run by the retail actor, and digital content can be easily and rapidly updated. This can create communication possibilities, for example, for real-time advertisement where the storage situation, weather conditions or other contextual data is used for automated content generation.

The reason for the low percentages might be that the content of pilot displays was conservative in nature. The effectiveness of in-store advertising largely depends on the content of the message; shoppers are most responsive to messages that relate to the task at hand and their current need state [1]. There are several means that might help us improve the effects of digital displays in the retail environment. Our data indicates that animation alone is not enough. In the future, we should extend our study to test stronger manipulations of message content and format. For example, stronger colors, contrasts and the use of audio could possibly attract the attention of store customers more effectively. In addition, the use of interactive elements could capture the customers' attention better. One future development idea would be also to enhance digital display systems with people tracking, e.g. by using depth sensors [6]. For example, the distance, amount, size and direction of the customers can be tracked and utilized when showing the changing content on the display.

Based on our results, it seems that younger generations see digital displays as being more beneficial in the retail environment than the older generations. A natural reason for this might be that younger people are used to collecting and utilizing digital information from a wide array of digital sources. Some terms and concepts from the digital era may also be unfamiliar to senior citizens. For example, our pilot study interviews revealed that it was not perfectly clear for some senior citizens what digital displays are; as a result, they were not prepared to pay attention to digital in-store information. Instead, older store customers largely reported that they still mostly utilize the ‘traditional’ media sources of advertising and communications, such as local newspapers and ad signs placed outdoors in front of the stores.

Behavioral measures of signage effectiveness such as shopper attention, product interest, and sales should be included in a future study. Furthermore, in the next phases of the research the effect of the physical positioning of displays in the store could be studied in more detail. Although we did not include questions related to display size or positioning, the topic came up spontaneously in various occasions. In our study all of the digital displays (except for one small display) were located relatively high in the store. Eight respondents proposed that they would be easier to see if they were on a lower level. The manager of the store where the small display was located had observed

that the display was very popular among the customers. Huang et al. [4] claim that small displays may encourage or invite prolonged viewing in public spaces to a greater extent than large displays. However, without further study we cannot say whether the customer preference for this display in our study was due to the small display size or, for example, the positioning of the display near the cash register queue, as suggested by some of our respondents. Positioning the display at eye-level has been found to be far more effective at attracting glances [4], and people prefer digital signage that is located where they could pause or wait [8]. The degree of engagement seems to vary depending on the shopper's angle of approach and proximity to the display screen [4]. We aim to address display size and positioning, as well as other issues discussed above, in our future work.

## References

1. Burke, R.: Behavioural effects of digital signage. *J. Advertising Res.* **49**, 180–185 (2009)
2. Chaffey, D.: Display advertising clickthrough rates (2013). <http://www.smartsights.com/internet-advertising/internet-advertising-analytics/display-advertising-clickthrough-rates/>. Accessed on 26 November 2014
3. Höffler, T., Leutner, D.: Instructional animation versus static pictures: A meta-analysis. *Learn. Instr.* **17**(6), 722–738 (2007)
4. Huang, E.M., Koster, A., Borchers, J.: Overcoming assumptions and uncovering practices: when does the public really look at public displays? In: Indulska, J., Patterson, D.J., Rodden, T., Ott, M. (eds.) PERVASIVE 2008. LNCS, vol. 5013, pp. 228–243. Springer, Heidelberg (2008)
5. Keränen, V., Lamberg, N., Penttinen, J.: Digitaalinen Media (in Finnish). WS Bookwell, Finland (2005)
6. Mäkelä, S.-M., Sarjanoja, E.-M., Keränen, T.: Treasure hunt with intelligent luminaires. In: Proceedings of AcademicMindTrek 2013, pp. 269–272. ACM (2013)
7. Müller, J., Wilmsmann, D., Exeler, J., Buzeczk, M., Schmidt, A., Jay, T., Krüger, A.: Display blindness: the effect of expectations on attention towards digital signage. In: Tokuda, H., Beigl, M., Friday, A., Brush, A., Tobe, Y. (eds.) Pervasive 2009. LNCS, vol. 5538, pp. 1–8. Springer, Heidelberg (2009)
8. Newman, A., Dennis, C., Wright, L.-T., King, T.: Shoppers' experiences of digital signage – A cross-national qualitative study. *Int. J. Digit. Content Technol. Appl.* **4**(7), 50–57 (2010). AICIT, Korea (Rep. of)
9. Nielsen, J.: Banner blindness: Old and new findings (2007). <http://www.nngroup.com/articles/banner-blindness-old-and-new-findings/>. Accessed on 26 November 2014
10. Ravnik, R., Solina, F.: Audience measurement of digital signage: Quantitative study in real-world environment using computer vision. *Interact. Comput.* **25**(3), 218–228 (2013)
11. Stern, A.: 8 ways to improve your click-through rate (2010). [www.imediaconnection.com/content/25781.asp](http://www.imediaconnection.com/content/25781.asp). Accessed on 26 November 2014

# Toward a Deeper Understanding of Data Analysis, Sensemaking, and Signature Discovery

Sheriff Jolaoso<sup>1</sup>, Russ Burtner<sup>2</sup>, and Alex Endert<sup>3()</sup>

<sup>1</sup> Virginia Tech, Blacksburg, USA  
[sheriff1@vt.edu](mailto:sheriff1@vt.edu)

<sup>2</sup> Pacific Northwest National Laboratory, Richland, USA

<sup>3</sup> Georgia Tech, Atlanta, GA, USA  
[endert@gatech.edu](mailto:endert@gatech.edu)

**Abstract.** Data analysts are tasked with the challenge of transforming an abundance of data into knowledge and insights. This complex cognitive process has been studied, and models created to describe how the process works in specific domains. Two popular models used for this generalization are the sensemaking and signature discovery models, which apply a cognitive and computational focus to describe the analytic process, respectively. This work seeks to deepen our understanding of the data analysis process in light of these two models. We present the results of interviews and observations of analysts and scientists in four domains (Biology, Cyber Security, Intelligence Analysis, and Data Science). Our results indicate that specific aspects of both models are exhibited in the analysts from our study, but neither describe the holistic analysis process.

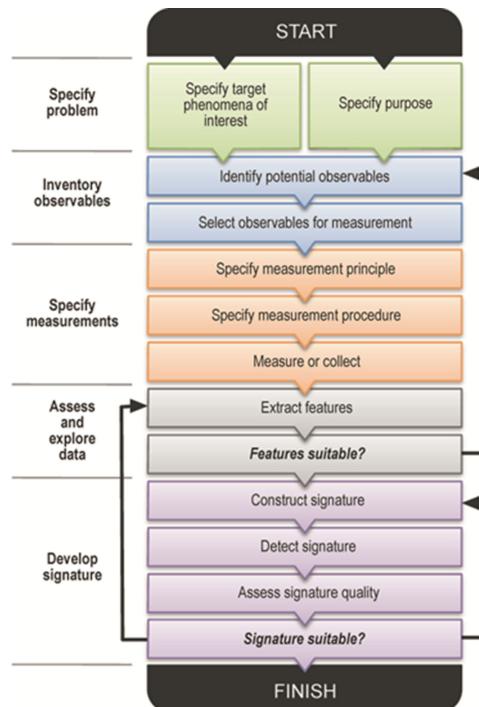
**Keywords:** Analytic process · Sensemaking · Signature discovery · Visual analytics · Data analysis

## 1 Introduction

Data analysts are tasked with coming about insightful conclusions based on large amounts of unstructured information provided to them. As such, understanding and generalizing the process of gaining such insight has been widely studied. For example, the literature on *sensemaking* depicts the iterative process of developing insight through foraging information and synthesizing knowledge by mapping features from data to a sets of hypotheses, and then performing assessments to understand the resulting complex relationships [1]. Sensemaking is a cognitive activity that entails the iterative development, re-assessment, and refinement of knowledge structures formed by prior experiences and data. Two popular models of sensemaking, Pirolli and Card's sensemaking loop [2] and the data/frame model presented by Klein et al. [3], present frameworks by which this process can be thought of. The emphasis of both of these models is on the cognitive aspects of users performing such tasks.

Another model, called the *signature discovery process* [4], is a model aimed at understanding the data-centric counterpart to such tasks. This approach (shown in Fig. 1) takes a more quantitative and computational approach to describing the analysts'

processes. Signature discovery describes the process of developing a signature based on an insightful understanding of the data gathered from both computational and cognitive reasoning. The signature discovery process requires analysts to specify a problem, and categorize the necessary data that needs to be inventoried, assessed, and explored. From this data, analysts can find a unique or distinguishing measurement, pattern, or collection of the data that can be used to detect, characterize, or predict a target phenomenon (e.g., state, object, action, or behavior) of interest. As a result, the *signature* represents a mathematical quantification of the insight. For example, a signature can be realized as a parameterized classifier used to generate a cluster of images, where that cluster represents an interesting finding as determined by the user.



**Fig. 1.** Diagram of the signature discovery process (adapted from [1]).

In this paper, we present the findings of a user study consisting of semi-structured interviews and observations of data analysts in four domains: Biology, Cyber Security, Intelligence Analysis, Data Science. For this paper, we broadly define *data analyst* and *analyst* to describe a person who is tasked with gaining insights from data, not as a specific job title or formal description of their task. The study seeks to understand how the practice of data analysis across these four domains compares to two current models describing such processes: the sensemaking loop and signature discovery. Our results indicate that while there exists overlap with current models,

as the fluidity and personalization of the participants' workflows carries inherent complexity. The primary contributions of our work are:

- Providing a deeper understanding of the data analysis processes in the four domains studied (Cyber Security, Intelligence Analysis, Biology, Data Science)
- Discovering that both sensemaking and signature discovery models depict specific aspects of data analysis, but neither fully describe the holistic process
- Discussion of findings to inform design of future data analysis tools.

## 2 Related Work

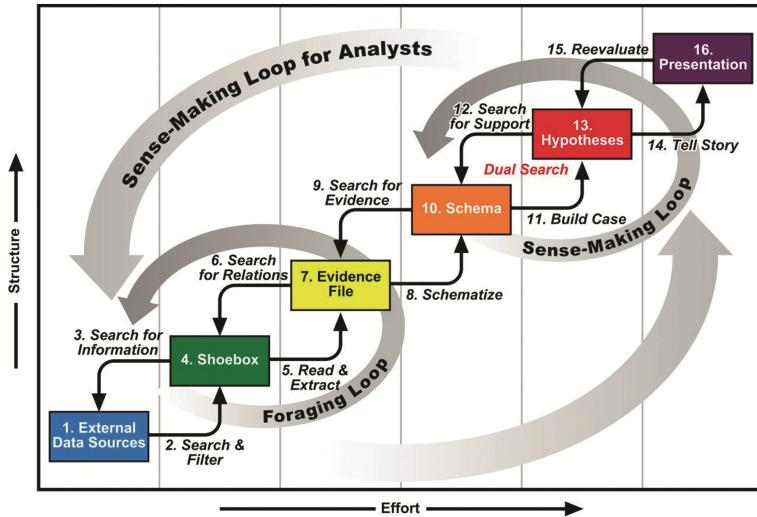
### 2.1 Signature Discovery

Baker et al. define a signature to be a unique or distinguishing measurement, pattern, or collection of data that detects, characterizes, or predicts a target phenomenon (state, object, action, or behavior) of interest as well as the transformation from events to measurements to features to categorical labels and associated uncertainties [4]. The steps of the signature discovery process are diagrammed in Fig. 1. The signature discovery process is a linear process that entails problem specification, identification of observable data, specification of measurements for the observable data, feature extraction, and finally signature development with the option to re-assess specific steps. The result of signature discovery, along with a signature, is a signature system; a reusable function, or piece of insight that can be used on similar inputted datasets. Once applied, the process calls for users to test the suitability of the created signature for the task and domain [4].

This work seeks to further understand the definition of signatures, signature systems, and the signature discovery process from the viewpoints of analysts from the four domains chosen. Further, these interviews and observations will help define the relationship between signature discovery and other sensemaking literature.

### 2.2 Sensemaking

Sensemaking, as defined by Russell et al., is the process of searching for a representation to answer task-specific questions [1]. There are two major loops of activities in sensemaking – a foraging loop that involves processes aimed at seeking information, searching and filtering it, and reading and extracting information and secondly, a sensemaking (or synthesis) loop that involves iterative development of a mental model from the schema that best fits the evidence [1, 2, 5]. A model of the sensemaking (the sensemaking loop) is shown in Fig. 2. The foraging loop progresses analysts from external, raw data sources to evidence files that refer to a more case-specific subset of information. The synthesis portion of this loop describes the steps involved in going from an evidence file to a final presentation of the hypothesis developed through schematizing, building a case, searching for evidence and support, and re-evaluating output, making this loop more of a cognitive, synthesizing activity. The emphasis of sensemaking is on creating a mental model and does not explicitly contain components correlating to scientific tools, such as visualizations [6].



**Fig. 2.** The sensemaking loop, presented by Pirolli and Card [5], depicting cognitive stages of data analysis.

Another model of sensemaking, presented by Klein et al. [3], shows another descriptive model for sensemaking. Their model depicts an exchange of information between the human and the data in terms of frames. That is, a user has an internal “frame” that represents her current understanding of the world. Then, as she continues to explore a particular dataset, her frames of the world are mapped against the information she uncovers. If the information supports a specific frame, that frame is thought to strengthen. However, when evidence is discovered through exploration that refutes the existence of such a mental frame, the frame can either be augmented or a new one created. Their data/frame model consists of reframing data through filtering, making connections, and defining characteristics, which can be equated to the foraging loop from Pirolli and Card’s work. Other steps and cycles elaborate frames, refine frames through questioning and preserve frames, and can be said to parallel aspects of the sensemaking loop defined in Pirolli and Card’s work.

## 2.3 Workflow Management Tools

It is assumed that there are computational steps in the analytic process. To carry out these computational steps, workflow management tools have been developed. It is expected that during the analytic process, that workflow management tools be used to automate computationally attractive steps. The steps used within these tools can potentially map to steps in the sensemaking and signature discovery processes.

For example, Taverna is a workflow management tool, first used in the area of bioinformatics, to assist in the procedural computation steps of conducting experiments [7]. The tool also provides provenance information, which is information that identifies the source and processing of data by recording the metadata and intermediate results

associated with the workflow. The provenance tracking capability enables users to conduct sensemaking activities, specifically those found in the foraging loop and the beginning of the sensemaking loop, through the use of visualizations and a graphical user interface. Steps of the signature discovery process where data is processed, such as feature extraction and measurement specification, can be paralleled to the calculation and data processing steps that a workflow management tool performs.

Similar, VisTrails visualizes the steps and interactions of the analytic process [8]. The tool captures the provenance of the visualization processes and the data they manipulate as well, which can enable the reevaluation steps within the sensemaking model. The tool also provides the ability to apply a dataflow instance to different data, which can be equated to the feature extraction aspect of signature discovery.

In this work, the importance of these tools and how they coincide with the signature discovery process and sensemaking is investigated.

## 2.4 Observing Analysts and Domain Experts

Analysis and assessments of the analytic process have been conducted to find out the challenges faced by analysts, design implications for analytic tool development, and future trends of the analytic process. For example, Fink et al. and Endert et al. observed how cyber analysts use visualizations and large high-resolution displays to perform a sensemaking task, producing findings about their processes [9–11]. The findings presented enhancements for analyst’s performance, such as desire to save the state of their investigation at certain points, a need to provide rich linkages among multiple visualization tools, means of keeping a visual history of steps in the analysis, and deep access to the data provided. Similarly, Kandel et al. interviewed analysts to gain insight on their analytic process within the social and organizational context of companies. They were able to create archetypal patterns to categorize analysts and organizational schemes of workflow processes for enterprise workers [12]. Evaluations of analytic processes were also investigated by Scholtz et al., who (through the use of the VAST challenge) found the importance of assessing the process of the analysis separate from the accuracy of the resulting analysis [13].

The work presented in this paper builds upon the knowledge gained through these studies. First, we study additional domains to gain insights into how data analysis occurs across domains. Second, we map our findings to two existing models for data analysis (signature discovery and sensemaking).

## 3 Method

The study consisted of a series of semi-structured interviews and observations of participants performing analysis to illuminate domain-specific information from our subject matter experts. The questions asked during the interviews include what type of problems they are tasked with, what types of data they use, which tools they use in their analytic process, and how they measure success. The observations were conducted by having the investigators passively watch as the analysts walked through one of their recent tasks.

The sporadic (and at time chaotic and time-sensitive) nature of their actual analysis forced us to observe past data analysis processes.

### 3.1 Research Questions

We seek to address the following three research questions (RQ1 – RQ3), and their associated hypotheses (H1 – H3):

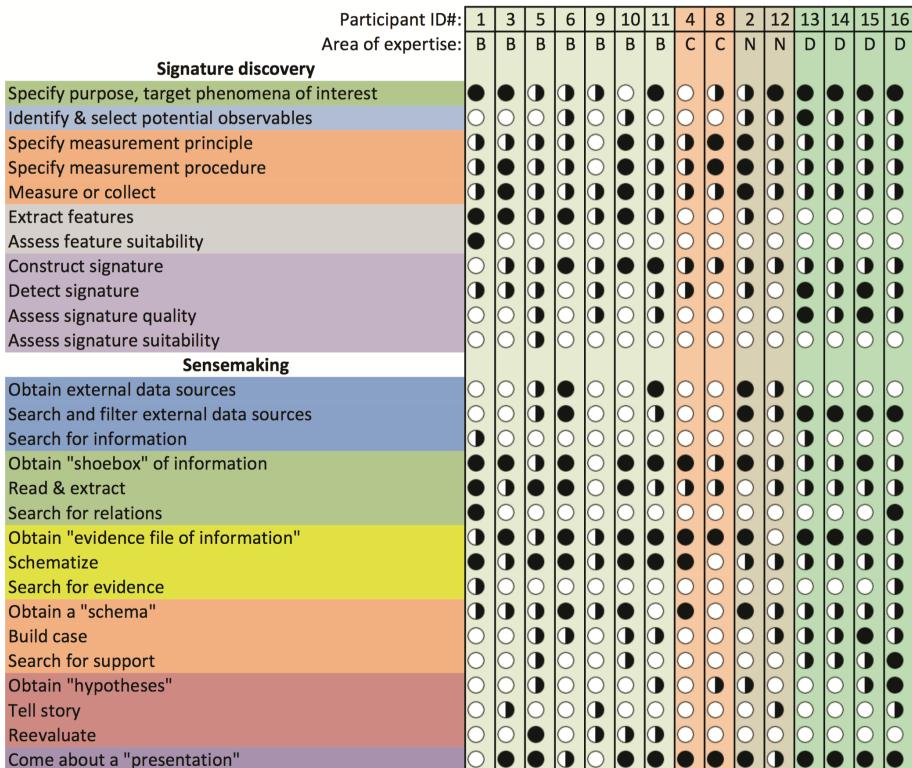
- RQ1:** What are the similarities and differences between the signature discovery process from Baker et al., the model of sensemaking from Pirolli and Card, and the analytic process as described by analysts?
- H1:** Our hypothesis for RQ1 is that the signature discovery process will reflect a more quantitative and sequential process, whereas sensemaking will be shown to be more fluid (as described in the related work). It is likely that the two may complement each other.
- RQ2:** What are tangible outcomes/products of the analytic process?
- H2:** We hypothesize that the outcome of the analytic processes is a signature as defined in [1] with little variance between domains and that a combination of the signature and signature system will form the analyst's presentation. That is, analysts seek to find repeatable mathematical descriptions of their findings.
- RQ3:** How do analysts use workflow management tools in their analytic process?
- H3:** Our hypothesis for RQ3 is that analysts use workflow management tools within their analytic process for feature extraction and signature construction (i.e., the more computationally-focused aspects of their work).

### 3.2 Data Analysis

From the semi-structured interviews, the investigators collected answers to the seed questions, captured on paper in natural language. These were later transcribed to digital versions, to ease indexing, searching, and annotating. The observations were conducted by having the investigators passively watch as the analysts walked through one of their recent tasks. The sporadic (and at time chaotic and time-sensitive) nature of their actual analysis forced us to observe past data analysis processes (instead of active analysis tasks). We collected notes on general observations, steps taken, and comments made by the analysts during these sessions. These notes were also transcribed to digital versions, similar to the interview responses.

Each of the three investigators performed an open coding process of binning the digital transcripts to reveal higher-level themes. While there were no formal, pre-defined topics or themes to code for, the investigators were aware of the previous models being tested (signature discovery and sensemaking). Thus, as each investigator individually analyzed the transcripts, caution was taken to mark instances where the transcripts describe or comment on aspects of either sensemaking or signature discovery. After each individual investigator saturated the individual themes (or bins), we met collaboratively to compare, contrast, and final set of labels for observed instances. These were

then mapped into the stages of the sensemaking and signature discovery models, shown in Fig. 3. Note that this set does not consist of all the stages of sensemaking and signature discovery, as some stages or aspects were not observed.



**Fig. 3.** Summary of how the analytic processes of analysts interviewed relate to the signature discovery and sensemaking processes based on our data analysis. Full circles indicate steps explicitly mentioned or observed. Half circles indicate steps that were generally mentioned, but not explicitly performed. Empty circles indicate steps that were not observed or mentioned during the interviews. The areas of expertise are N = Intelligence Analyst, B = Biology, C = Cyber Security, and D = Data Science. P7 was removed for data classification and sensitivity concerns.

### 3.3 Participants

Semi-structured interviews were conducted with sixteen analysts (P1–P16) in the areas of Biology (7), Cyber Security (3), Intelligence Analysis (2) and Data Science (4). Of these sixteen, one had to be discarded for analysis due to classification and data sensitivity concerns (P7).

**Biology.** The biologists that we interviewed covered a wide scope of work from Epidemiology to Genetics. Six of the seven hold a Ph.D. in Biology/Computational

Biology, and have five or more years of experience in the field with strong publishing credentials.

**Cyber Security.** The cyber security analysts had strong IT backgrounds and have been established in the field for 10 or more years (two were only in their current job for 2 years). Their scope of work covered smaller local networks to broader more national investigative cyber networks. Their levels of education are Bachelors of Science to Masters in Computer Science or Computer Engineering.

**Intelligence Analysts.** The intelligence analysts we interviewed worked for multiple government agencies including the Department of Energy. Each had a diverse background in either a nuclear science related field (thus the “Nuke” label in figures in this paper) or computational science. Both have been in intelligence for more than ten years.

**Data Science.** The data scientist we talked to are all traditional computer science or formal statisticians. Their level of expertise varied from over 10+ years of experience to only two years out of school. Their level of education also varied from Master’s degree to Bachelors of Science.

## 4 Findings

Through the thematic coding approach used to analyze the data collected from the interviews and data analysis observation sessions, we present the following findings. First, we present details regarding the characteristics of the analytic process based on the data sources, tools used, and methods for recording their analytic provenance (or process). Second, we compare the self-reported and observed processes to existing models of sensemaking and signature discovery. Third, we analyze the analysts’ methods for reporting the findings and insights.

### 4.1 Analytic Process Characteristics

**Data Sources.** All of the domains we studied utilized databases as a primary method for storing and retrieving their data. Among the intelligence analysts, data sources commonly used were media such as images and text. Our biology users accessed many scientific databases shared among other scientists in their discipline or clinical study. The Data Science analysts were mostly users of a variety of observational sensor data gathered from databases, log files, or data streamed directly from a sensor or system. These datasets were described to be rather large and high dimensional. The cyber security analysts primarily accessed locally-hosted databases of log files. One challenge faced by all of our analysts we spoke with was how to subset the large quantity of data, so that a smaller amount of it can be worked on locally.

**Tools Used.** Most of the analysts across the four domains used a data analytic tool, like R or Matlab, as well as a variety of visualization tools. A commonality among the analysts across domains was the use of spatialization tools, which are visualizations that

cluster like data together based on specific attributes of the data. Spatialization tools such as Starlight [14], IN-SPIRE [15], Clique and Traffic Circle [16], and ARGIS were used by the analysts that we interviewed. These spatialization tools were semi-autonomous processes that allowed manipulations by the analysts to occur in order to perform feature extraction. A myriad of custom scripts were also used by analysts as well as other feature extracting processes, which were used in a more batch/sequential manner in comparison to the spatialization tools focused on exploration.

To perform any of their scripts and queries, it was assumed that analysts utilized workflow management tools, such as Taverna, but that did not prove to be true among the majority of analysts (only two stated they use such tools). For example, P11 stated that Taverna “provided too many options I do not need.” This analyst preferred to maintain a detailed record of inputs and outputs of scripts and queries to recall the process and execute the sequence of scripts for the task. P9 detailed that the purpose of using the Galaxy workflow management tools was to “replay workflows” as well as “share and collaborate on workflows”, something that was not desired and thus added unneeded overhead to using the tool. Other participants detailed natural progressions in their work, which may have been complimented by a workflow management tools. While they commented that these sequences of scripts could be described as a workflow, they are created *in situ*, and thus they prefer the flexibility of performing these computational steps as scripts, often executed in standard command line interfaces. These findings were particularly interesting with regards to our third research questions and hypothesis (R3 and H3), in that workflow management software, in the computational sense, was not used by our participants.

**Analytic Provenance Recording and Documentation.** Upon coming across phenomena of interest, analysts look to capture/document it in some way in order to recall it later. These moments are captured in two ways: the process and the finding. P11 captures the process by using a “configuration file to specify what to run”, so that it can be repeated when needed. P2, an analyst that performed most work manually, filed pages (printed hard copies) to capture findings. P3 stated that findings are best captured and described through publications. From the interviews, it was seen that there is an importance set on documenting and filing phenomenon for sharing and revisiting findings. P15 stated that he and any collaborators kept an online journal that was used to keep track of important findings throughout their processes for easy reference later on and to provide an opportunity for peer reviewing. In general, we observed that the provenance or process of our analysts was maintained primarily via inputs and outputs of data to and from scripts.

## 4.2 Comparisons to Existing Models

Our hypothesis (H1) for comparing the observed processes with both sensemaking and signature discovery is that signature discovery would be more structured (even, sequential), compared to a more fluid sensemaking process. Below, we describe our findings in more detail.

**Comparison to the Signature Discovery Process.** Figure 1 describes the signature discovery process presented by Baker et al. [4]. Through the interviews, we found that analysts did not follow the linear design of this model. One noticeable difference between the analytic process described by participants and the model of the signature discovery process was the exclusion of certain steps that are present within the signature discovery model. For example, the majority of analysts interviewed did not take inventory of observables; typically the step where data intended for use is defined. In the case of most analysts, the data provided to them for analysis did not require a definition or schema to define the atomic units in the data. Another difference was the lack of assessment and exploration of data, mainly seen amongst the intelligence analysts interviewed. P11 stated that there were “no feature detection models” needed, as the data was structured upon reception by the analyst. P2 and P12 noted how the feature extraction process was not computational and expressed the desire to have “entity extraction in the future.” We also saw that there was a lack of explicit description of assessments, detail in the discussion section.

The participants also emphasized the cyclic nature of the analytic process. P5 described the process as the “refinement of a filter to understand the tradeoffs in the data”. This cyclic process follows the loop between signature suitability and extracting features within the signature discovery process.

Within the Biology domain, we saw that all analysts performed most of the data-intensive tasks within the signature discovery process, mainly feature extraction, and measuring, and collecting data. We found that the Cyber and Data Science analysts did not perform feature extraction for the most part because the data provided to them was typically already structured and mostly quantitative data, so their main concern was to gain insight from what was provided. This is a similar case for the Intelligence analysts. However, Intelligence analysts differ from the other domains in that they identified observables as a critical part in their process. Analysts in all domains came to an endpoint or what could be called a signature, based on the data provided to them. There were a variety of endpoints that came about, such as patterns to be compared to databases, feature sets, and points of interest.

There are two forms of outcomes described by the analysts we interviewed that can be described in terms of the signature discovery process. One is a signature, which is a unique or distinguishing measurement, pattern, or collection of data that detects, characterizes, or predicts a target phenomenon of interest. In the case of the Cyber domain, this echoes prior work by Fink et al. that found the final produce of cyber security analysts to be a complex query [10]. Another form of result is a signature system, which describes a sequence of steps taken to attain the insight. For example, P6 stated how the findings attained through the use of scripts, filters, and other data transformations are “not valuable without remembering how all of these steps fit together”.

**Comparison to the Sensemaking Loop.** There are differences based on the sensemaking loop presented by Pirolli and Card [5], and the processes observed. One difference that was persistent throughout most of the participants was the lack of top-down reassessment steps performed.

Another component of the sensemaking process that was not present in the analytic process of most of our analysts was the steps prior to obtaining a shoebox of evidence. Most analysts were provided scoped, domain-specific data so there was no need for searching and filtering external, unrelated information. Most analysts did work to approach a presentation as a deliverable in order to find if the analytic process should be continued or if the final result had been sufficiently reached. Kang and Stasko's findings are similar to these findings, in that they identified different categorizations of analyses: structured and intuitive [17].

For the Cyber and Data science domain, we found the synthesis stages often lacking. In the Cyber domain, based on their described analytic techniques, there is less information synthesis. Their analytic processes were more focused on foraging. P8 commented that the "findings of our queries are sent off and evaluated" by other collaborators. Similarly, the participants from the data science domain commented that they were sometimes required to simply forage through data. That is, to transform data from one (or a set of) structured or unstructured sources to a single, clean, unified spreadsheet or database.

The biology and intelligence analysts, on the other hand, performed more of their work in the synthesis part of the sensemaking loop, specifically schematizing. The biologist exhibited this in the form of generating research questions and hypotheses. The intelligence analysts often created hypothetical scenarios and observe how they panned out with respect to the current and new data.

### 4.3 Products and Results of the Analytic Process

We analyzed the data collected from the interviews and observations to determine what constituted as "products" from the analyses of our participants. These are primarily the *insights* our participants found, and also the *repeatable processes* that allowed them to acquire the insights. The details below help address our second hypothesis (H2), regarding the outcomes and products.

**Insights.** A majority of the analysts consider a valuable outcome of their process to be broadly categorized as novel *insight*. P10 stated that a goal is to "represent data in a meaningful way" and that this is often through the detection of a "pattern found in the data." P8 stated that a valuable finding was the "unique identification for features," which can be considered a derived facet from the original data to gain insight. P4 stated that findings typically come in the form of behavior of extracted features (biomarkers). P9 defined a the valuable insights as a "piece of info that stands out from background" and "that gives you meaning."

We found that the insights of the analysts in our study were also focused on features or characteristics in the data, rather than the data objects themselves. For example, P3 states that characteristics of "a panel of genes", such as "how it behaves", "a set of features/markers", and "abundance level[s]", are valuable results of analysis. This form of an analytic product maps well onto the prior work of defining what a signature is in different contexts. For example, signatures have been described as nucleotide sequences [18, 19], features of network communication packets [20], or a set of proteins and their measured abundance levels [21].

**Repeatable Processes.** Analysts also described their findings in terms of the process (rather than the outcome). P5 stated that findings are typically a “sequence of up-front data processing, followed by decision making unique to the data” to come about a set of answers. These steps and decisions made are the critical findings, as they help inform collaborators of how the findings were generated. P5 also described the process as “hypothesis generation, not hypothesis answering.” P11 stated that 75 percent of the process was processing data through mass spectrometry and matching it to a database, and “the other 25 percent is generating p-values”.

#### 4.4 Collaboration

The analysts exhibited variations of informal collaboration during their analysis. P6 often performs “pair analytics” [22], in which P6 and a domain expert work alongside an expert in the area they’re performing analysis for. P3 leverages expert knowledge in combination with “multiple sensor sources to find patterns” with the goal of providing “trained models to get experts further along … to make better distinctions in features.” However, there was no explicit collaboration between analysts observed or reported (with the exception of handing off results and steps in a sequential manner to collaborators). For example, the “identify and select potential observables” step of the signature discovery process was not performed because another analyst had previously completed this step and communicated the resulting data to the next analyst.

### 5 Discussion

#### 5.1 Sensemaking and Signature Discovery

The entirety of the analytic process can be said to be composed of aspects of both sensemaking and signature discovery. Figure 3 shows a summary of how the analytic processes of analysts interviewed relate to the signature discovery and sensemaking processes based on our data analysis. Full circles indicate steps explicitly mentioned or observed. Half circles indicate steps that were generally mentioned, but not explicitly performed. Empty circles indicate steps that were not observed or mentioned during the interviews. As can be seen by these figures, it is rare for individual analysts to conduct every step explicitly.

The signature discovery process is computationally based procedure. From the interviews, it was shown that data processing was a necessary part of the analytic process. The goal of signature discovery is quantitative and well defined, where a signature and signature system are the ideal results of the process. In contrast, sensemaking is primarily a cognitive process and is not as rigidly structured as the signature discovery process. There are more avenues for assessing previously performed tasks in the process compared to signature discovery. As a result, we saw similarities with the sensemaking process in the more synthesis-focused analysis tasks, and more similarities to signature discovery when data-centric methods were used.

Further, while specific steps in the foraging loop map directly to some of the steps in the signature discovery process, a link between the synthesis stages from sensemaking

to signature discovery are not as apparent. For example, the task of transitioning from external data sources to an evidence file maps to the steps in the signature discovery process steps of taking inventory of observables to feature extraction. From there, the two processes appear to branch off into independently.

As a result, we found that both models do not encompass the process of analysis across all of the domains fully. The steps they performed were mostly based on the data sources and tasks given. Signature discovery and sensemaking have been identified as independent processes, focused on the computational and cognitive aspects of data analysis, respectively. The differences are also illuminated in observing the result of the models. Signature discovery has the signature as a result, while sensemaking suggest a presentation to share or present findings, processes, and knowledge to someone else. From our study, we find that the distinction between these two is dependent primarily on the domain. In Cyber Security, for example, a repeatable query or classifier is often desired, whereas intelligence analysis emphasizes the shared knowledge (i.e., the intelligence being discovered) to give to someone else.

## 5.2 Hypotheses and Assessments

The formation and assessment of hypotheses played a central role in the analytic processes of our participants. The position of the hypothesis in the analytic process can vary, depending on the initial starting point of the analysis. Proving or disproving a hypothesis can be the driving force behind the analytic process, or developing a hypothesis can be the purpose of the analysis. It was found that some of the analysts' processes were based on either hypothesis generation or hypothesis verification that has an impact on how the analytic process is conducted. P1 stated the measure of success for their work was "hypothesis confirmation and verifying a prediction," which correlates to the scientific method. P5 stated that analysis was about "hypothesis generation" and P2 stated that the goal for their work is a hypothesis. Among the data scientists interviewed, the placement of hypothesis formation within their analytic processes significantly varied depending on the question that needed to be asked of the data and the type data being dealt with. Typically, when dealing with cyber data, the Data analysts had a hypothesis near the earlier steps of their process with an end goal of verifying their hypothesis. Hypothesis generation can also arise when there is no specific initial problem. In this case, the analytic process is used to develop more refined problems as opposed to solving a succinct problem or question.

Our participants also exhibited tactics of hypothesis assessment. As seen in Fig. 3, many analysts do not explicitly state that they conduct assessments of suitability in signature discovery, or any of the top-down analytical steps within the sensemaking loop. However, during the observation of analysis, it appears these steps are inherent to the analytic process that analysts do not consider them to independently coincide with the signature discovery or sensemaking process. This is reflected in Pirolli and Card's "dual search" activity in the sensemaking model [5], and may evidence the importance of that aspect of the sensemaking process. Similarly, the analytic processes observed can be generally fit either the "top-down" or "bottom-up" terminology [2, 5], as hypothesis-driven analysis, or the process of forming hypotheses.

### 5.3 Implications for Design

The results of this study can help inform the design and implementation of systems created to support data analysis. Through gaining a deeper understanding of the cognitive and computational processes involved in gaining insights in data-driven domains, we can develop design guidelines and considerations.

There is an inherent importance of **capturing and tracking the analytic provenance** [23]. Analytic provenance encompasses the interactive steps, intermediate results, and hypotheses considered throughout analysis. The participants of our study commented that recalling their process after a focused data analysis session is difficult. Analytic provenance support in such tools needs to strike a fine balance between allowing users to maintain this “cognitive zone” [24] during analysis, and encouraging users to record and annotate their process. Our participants commented that such provenance support should not interfere with the actual data analysis process. Thus, there is an open challenge for the field to determine how formal and explicit such capturing and tracking techniques should be. In comparison, passive capture and interpretation of user interaction may be a more well-suited approach [25, 26].

Further, the analysts from the domains we studied emphasized the **importance of hypotheses during their analytic processes**. This included generation, testing, and validation of hypotheses. However, hypotheses were not typically formal (or mathematically grounded). For example, they came in the form of “hunches” or “moments of interest” for following one’s curiosity about the data. This also came through in the observations, where the investigators often asked how the analyst decided to try one approach over another. Often, the analysts stated that they did not know, or that they had seen something like it before. This finding echoes the importance of researching and developing more formal methods of tracking hypotheses, but also of evaluating exploratory data analysis tools.

## 6 Conclusion

In this paper, we present the results of a user study of professional analysis in four domains (Biology, Cyber Security, Intelligence Analysis, and Data Science). We analyzed the tasks and processes of analysts to get a deeper understanding of the cognitive and computational aspects of data analysis. In order to do this, we conducted a series of semi-structured interviews and analysis observations.

Our results indicate that the data analysis processes exhibited by our study participants was a combination of the known sensemaking and signature discovery models. Generally, the cognitively-focused aspects of synthesizing the information into knowledge was represented in the sensemaking models. In contrast, the computationally-focused data foraging and features extraction stages were better represented in the signature discovery model. The diversity of analytic processes, among those we interviewed suggests a combination of sensemaking and signature discovery as a viable model for data-driven analytic processes. These findings are further discussed, including implications for design, which can inform the design and implementation of future data analysis systems.

**Acknowledgements.** This research is part of the Signature Discovery Initiative at Pacific Northwest National Laboratory, conducted under the Laboratory Directed Research and Development Program at PNNL, a multi-program national laboratory operated by Battelle for the U.S. Department of Energy.

## References

1. Russell, D.M., Stefik, M.J., Pirolli, P., Card, S.K.: The cost structure of sensemaking. In: Proceedings of the INTERACT 1993 and CHI 1993 Conference on Human Factors in Computing Systems, New York, NY, USA, pp. 269–276 (1993)
2. Pirolli, P., Card, S.: Sensemaking processes of intelligence analysts and possible leverage points as identified through cognitive task analysis. In: Proceedings of the 2005 International Conference Intelligence Analysis, McLean Va, p. 6 (2005)
3. Klein, G., Moon, B., Hoffman, R.R.: Making sense of sensemaking 2: a macrocognitive model. *IEEE Intell. Syst.* **21**(5), 88–92 (2006)
4. Baker, N., Barr, J., Bonheyo, G., Joslyn, C., Krishnaswam, K., Oxley, M., Quadrel, R., Sego, L., Tardiff, M., Wynee, A.: Research towards a systematic signature discovery process. In: IEEE Intelligence and Security Informatics Signature Discovery Workshop (2013)
5. Pirolli, P., Card, S.: Information foraging in information access environments. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, New York, NY, USA, pp. 51–58 (1995)
6. Pohl, M., Smuc, M., Mayr, E.: The user puzzle: explaining the interaction with visual analytics systems. *IEEE Trans. Vis. Comput. Graph.* **18**(12), 2908–2916 (2012)
7. Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., Carver, T., Glover, K., Pocock, M.R., Wipat, A., Li, P.: Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* **20**(17), 3045–3054 (2004)
8. Callahan, S.P., Freire, J., Santos, E., Scheidegger, C.E., Silva, C.T., Vo, H.T.: VisTrails: visualization meets data management. In: Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data, New York, NY, USA, pp. 745–747 (2006)
9. Endert, A., Andrews, C., Fink, G.A., North, C.: Professional analysts using a large, high-resolution display. Presented at the IEEE VAST Extended Abstract (2009)
10. Fink, G., North, C., Endert, A., Rose, S.: Visualizing cyber security: usable workspaces. *VizSec* (2009)
11. Andrews, C., Endert, A., North, C.: Space to think: large high-resolution displays for sensemaking. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 55–64 (2010)
12. Kandel, S., Paepcke, A., Hellerstein, J.M., Heer, J.: Enterprise data analysis and visualization: an interview study. *IEEE Trans. Vis. Comput. Graph.* **18**(12), 2917–2926 (2012)
13. Scholtz, J., Plaisant, C., Whiting, M., Grinstein, G.: Evaluation of visual analytics environments: The road to the Visual Analytics Science and Technology challenge evaluation methodology. *Inf. Vis.*, Jun. 2013
14. Risch, J.S., Rex, D.B., Dowson, S.T., Walters, T.B., May, R.A., Moon, B.D.: The STARLIGHT information visualization system. In: Proceedings of the IEEE Conference on Information Visualisation, Washington, DC, USA, pp. 42–49 (1997)
15. Wise, J.A., Thomas, J.J., Pennock, K., Lantrip, D., Pottier, M., Schur, A., Crow, V.: Visualizing the non-visual: spatial analysis and interaction with information for text documents. Presented at the Readings in Information Visualization: Using Vision to Think, 300791, pp. 442–450 (1999)

16. Abdullah, K., Lee, C.P., Conti, G., Copeland, J.A.: Visualizing network data for intrusion detection. In: Proceedings from the Sixth Annual IEEE SMC Information Assurance Workshop, IAW 2005, pp. 100–108 (2005)
17. Kang, Y., Stasko, J.: Characterizing the intelligence analysis process: Informing visual analytics design through a longitudinal field study, pp. 21–30 (2011)
18. Vieira, J., Mendes, M.V., Albuquerque, P., Moradas-Ferreira, P., Tavares, F.: A novel approach for the identification of bacterial taxa-specific molecular markers. *Lett. Appl. Microbiol.* **44**(5), 506–512 (2007)
19. Phillippy, A.M., Mason, J.A., Ayanbule, K., Sommer, D.D., Taviani, E., Huq, A., Colwell, R.R., Knight, I.T., Salzberg, S.L.: Comprehensive DNA signature discovery and validation. *PLoS Comput. Biol.* **3**(5), e98 (2007)
20. Han, H., Lu, X.L., Lu, J., Bo, C., Yong, R.L.: Data mining aided signature discovery in network-based intrusion detection system. *SIGOPS Oper. Syst. Rev.* **36**(4), 7–13 (2002)
21. Bock, C., Coleman, M., Collins, B., Davis, J., Foulds, G., Gold, L., Greef, C., Heil, J., Heilig, J.S., Hicke, B., Nelson Hurst, M., Husar, G.M., Miller, D., Ostroff, R., Petach, H., Schneider, D., Vant-Hull, B., Waugh, S., Weiss, A., Wilcox, S.K., Zichi, D.: Photoaptamer arrays applied to multiplexed proteomic analysis. *Proteomics* **4**(3), 609–618 (2004)
22. Arias-Hernandez, R., Kaastra, L.T., Green, T.M., Fisher, B.: Pair analytics: capturing reasoning processes in collaborative visual analytics. Presented at the Hawaii International Conference on System Sciences, pp. 1–10 (2011)
23. North, C., Chang, R., Endert, A., Dou, W., May, R., Pike, B., Fink, G.: Analytic provenance: process+ interaction+ insight. In: CHI 2011 Extended Abstracts on Human Factors in Computing Systems, pp. 33–36 (2011)
24. Green, T.M., Ribarsky, W., Fisher, B.: Building and applying a human cognition model for visual analytics. *Inf. Vis.* **8**, 1–13 (2009)
25. Endert, A.: Semantic interaction for visual analytics: toward coupling cognition and computation. *IEEE Comput. Graph. Appl.* **34**(4), 8–15 (2014)
26. Brown, E.T., Ottley, A., Zhao, J., Lin, Q., Endert, A., Souvenir, R., Chang, R.: Finding waldo: learning about users from their interactions. *Trans. Vis. Comput. Graph* (2014)

# HCI Practices in the Nigerian Software Industry

Abiodun Ogunyemi<sup>1(✉)</sup>, David Lamas<sup>1</sup>,  
Emmanuel Rotimi Adagunodo<sup>2</sup>, and Isaias Barreto da Rosa<sup>3</sup>

<sup>1</sup> Institute of Informatics, Tallinn University,  
Narva Mnt 29, 10120 Tallinn, Estonia  
[{abnogn,david.lamas}@tlu.ee](mailto:{abnogn,david.lamas}@tlu.ee)

<sup>2</sup> Department of Computer Science and Engineering,  
Obafemi Awolowo University, Ile-Ife, Nigeria  
[eadagun@oauife.edu.ng](mailto:eadagun@oauife.edu.ng)

<sup>3</sup> ECOWAS Commission, Niger House, Area 1, Abuja, Nigeria  
[isaiasbr@gmail.com](mailto:isaiasbr@gmail.com)

**Abstract.** In this paper we explore the state of HCI practices in the Nigerian software industry. Our societies have evolved quickly into an information age, and the criticality of software and humans as components of socio-technical systems becomes more worthy to address. In Nigeria, the level of HCI practices is not yet known. We understand clearly, the role of software systems and services to strengthen information societies, and we decided to run a survey of the local software organizations. The results from the survey indicate some level of HCI awareness. Therefore, we conducted some semi-structured interviews in order to deepen our understanding of HCI practices in the industry. The results show there is a knowledge limit regarding HCI practices in the industry. We present a preliminary report of the results obtained from our studies of software organizations in Nigeria.

**Keywords:** HCI · Human-Centred design · Human-Centred software engineering · Usability · HCI education

## 1 Introduction

Although the field of human-computer interaction has been in existence for more than three decades, its spread has yet to be significant. Most of the spread has been in the developed countries and developing countries continue to lag behind [17].

Nigeria has an overwhelming population of 150 million people, which also accounts for approximately 20 % of Africa's population (situation in 2012) [8]. Recently, the Nigerian government introduced a cashless economic policy, which implies that economic transactions are to be conducted electronically. Furthermore, many businesses have been moved online, and e-commerce is becoming a major business trend in the country [2]. Thus, the role of software systems and services

cannot be ignored in Nigeria. Approximately 33 % of Nigeria residents are connected to the Internet in a country where approximately 67 % of the Internet users are mobile.<sup>1</sup> Nigeria is in the 133<sup>rd</sup> position in the recent World ICT development index.<sup>2</sup> Although it is the biggest Economy in Africa (primarily because it is the biggest oil producer in Africa), Nigeria still remains a developing nation. It can be envisaged as well what leading role Nigeria could play in promoting HCI in Africa in terms of the country's size and economy.

The story of HCI uptakes in developed and developing countries so far, might not differ, especially, when talking about certain practices such as usability engineering, user experience and Human-Centred Design (HCD). For example, Larusdottir, Haraldsdottir, and Mikkelsen, [13], ran a survey of the Icelandic software industry to determine how practitioners perceive the importance of usability and user involvements in software projects. The authors found that most of the companies use their own method regarding user involvement methods, and more than a third of the organizations surveyed, are skeptical regarding the importance of usability. Similarly, Ji and Yun, [12], conducted a survey of 184 Korean IT professionals and 90 User Interface/Usability practitioners, regarding User-Centred Design (UCD) and usability practices in the Korean IT industry. Their results show that awareness of UCD/Usability is high, but the reality of its application in projects is not fully realized.

The Nigerian software industry is still very young and in its formative stage. There are not yet regulations for the industry and most software companies use in-house methods [6, 18]. There are many small companies and only a few of these companies focus on custom developments [14]. Most custom developments are largely based on web applications. Very few of these organisations develop off-the-shelf software in which lower-level applications such as payroll, human resources management, educational, and accounting solutions are built from scratch as semi-packages and configured for various customers over time [18]. There are very few universities that offer an elementary course in HCI. Thus, there are very scarce sources to describe the state of HCI practices in Nigeria.

This paper describes and discusses the results from a recent field study regarding the state of HCI practices in Nigeria. To the best of our knowledge, a study such as this has not been conducted in Nigeria so far.

In the next section, we present our method. Next, we present the results, and finally, we discuss the results and describe future work that needs to be done.

## 2 Method

An online survey was deployed using the LimeSurvey open source tool. Respondents were targeted from indigenous software companies in Nigeria. The focus of the survey was on how Nigerian software practitioners conduct usability, user experience and

---

<sup>1</sup> UN e-Government survey - [http://unpan3.un.org/egovkb/Portals/egovkb/Documents/un/2014-Survey/E-Gov\\_Complete\\_Survey-2014.pdf](http://unpan3.un.org/egovkb/Portals/egovkb/Documents/un/2014-Survey/E-Gov_Complete_Survey-2014.pdf).

<sup>2</sup> ICT development index - [https://www.itu.int/en/ITU-D/Statistics/Documents/publications/mis2014/MIS2014\\_without\\_Annex\\_4.pdf](https://www.itu.int/en/ITU-D/Statistics/Documents/publications/mis2014/MIS2014_without_Annex_4.pdf).

human-centred design practices. The questionnaire was designed in such a way that some questions would only pop up, if an earlier answered question is related to the next one. Furthermore, there were questions, which allow respondents to choose multiple options and there were some, which allow single option.

Prior to the commencement of this study, the Institute of Software Practitioners of Nigeria (ISPON) was partnered with and a list comprising 50 indigenous software companies was obtained. A database search on relevant websites was made and through this search, a total of 95 companies were invited through e-mail to join the study. Sixty-seven companies participated in the survey, which gave us a response rate of 70 %. However, only 22 responses were useful for our analysis. Forty-five responses were rejected because they did not answer at least 75 % of the questionnaire [10].

In order to strengthen and exemplify the results of the survey, we conducted ten semi-structured interviews in three indigenous software companies. The three companies were randomly selected. One of the goals of the interviews was to deepen our understanding of the keys issues regarding HCI practices in Nigeria. This study was conducted between October 2014 and December 2014.

## 3 Results

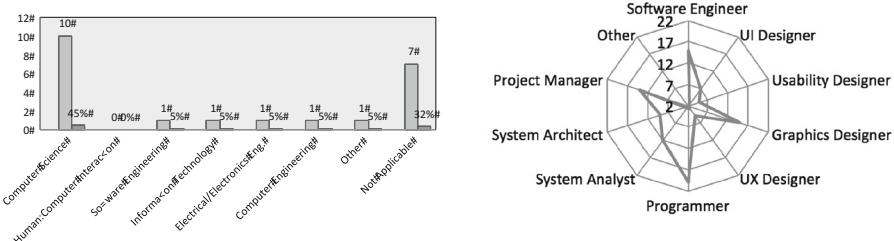
### 3.1 Demographics

A majority of the respondents came from the software development sector (10) and information technology services (9). Other sectors include telecommunication (1), non-governmental organisation (1) and e-commerce (1).

Regarding organisation size, 12 participants were from very small companies (10-20 staff), 5 were from small companies (50-90 staff), 3 were from medium companies (100-199 staff), and one participant each was from a large company (200-499 staff) and a very large company (more than 500 staff) respectively. This suggests that the Nigerian software industry is primarily composed of small companies. A previous study also suggested a similar trend [18]. Twenty companies are located in the South-West of Nigeria and one organisation each comes from the North and one from the South. Fifteen companies are located in Lagos. Lagos is the Economic Capital of Nigeria where the most prominent Nigerian Companies are located.

Regarding their educational qualifications, 15 respondents (68 %) possess a BSc degree, three respondents each (14 %), have diplomas and other certificates and one respondent (4 %), has a High School certificate. Figure 1 (left-hand side) is the overview of the respondents' background regarding their first degree. The results show that 45 % of the respondents possess a Bachelor's degree in Computer Science.

With respect to the composition of software teams, the results indicate that there are very few HCI experts in the teams, and that the major aspect of HCI found in practice is graphic design. The results are also in Fig. 1 (right-hand side). Regarding HCI awareness, 17 organisations (77 %) are aware and five organisations (23 %) are not aware.



**Fig. 1.** Respondents' first degree backgrounds and composition of software teams

Of the 15 respondents, who possess a BSc degree, only 9 indicate they took a course in HCI, and when asked to describe these HCI courses, some of the responses are: “*MYT364 - Fundamentals of interaction design*”

“*Covered just the basics of HCI*”

“*Visibility and Affordance*”

“*It involves the study, planning, design and uses of the interaction between people (users) and computers*”

The roles of the respondents in their organisations are Usability Designer (2), Programmer (7), Software Engineer (5), UX Designer (1), Project Manager (3), CEO (2), Technical Resource Engineer (1) and Chief Software Architect (1).

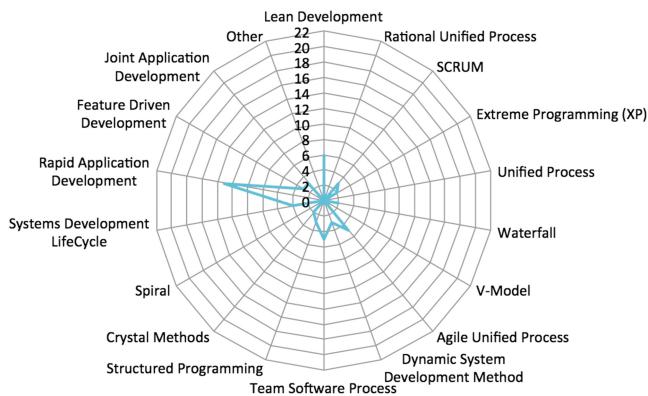
The responses regarding the HCI courses suggest that this category of practitioners received elementary HCI knowledge. However, it is imperative that these graduates are well-equipped with hands-on skills in order to succeed in the industry [15].

Regarding the years of experience of the respondents, the results obtained suggest that most of the respondents might be relatively young in their main roles. Eleven respondents (50 %), have less than five years of experience. Seven respondents (32 %), have 5–10 years of experience, one respondent (9 %) each had 10–15 years of experience and more than 15 years of experience respectively. In comparison with the respondents' educational backgrounds, it is possible that the bulk of the practitioners with less than five years of experience are those with some level of HCI education.

We wanted to know what software development methodologies are used in respondents' organisation. Figure 2 reveals that most of the organisations used the Rapid Application Development (RAD) methodology. Figure 2 also reveals that none of the software organisations used such methods as Rational Unified Process, Unified Process, V-Model and Spiral.

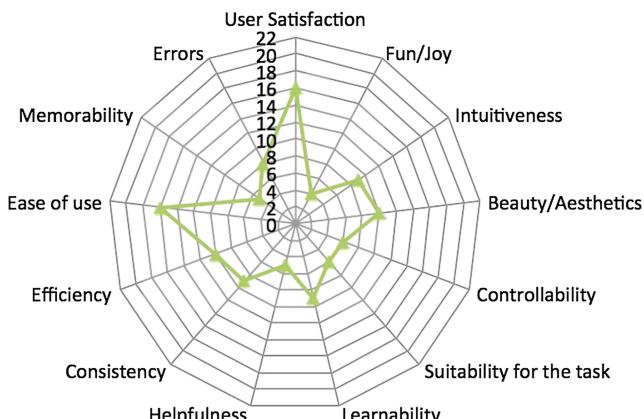
### 3.2 User Experience, Usability, and Human-Centred Design Practice

We asked if the organisations address user experience (UX). Seventeen organisations (77 %) were positive and five organisations (23 %) indicate they do not address UX.



**Fig. 2.** Software development methodologies used in respondents' organizations (multiple options allowed)

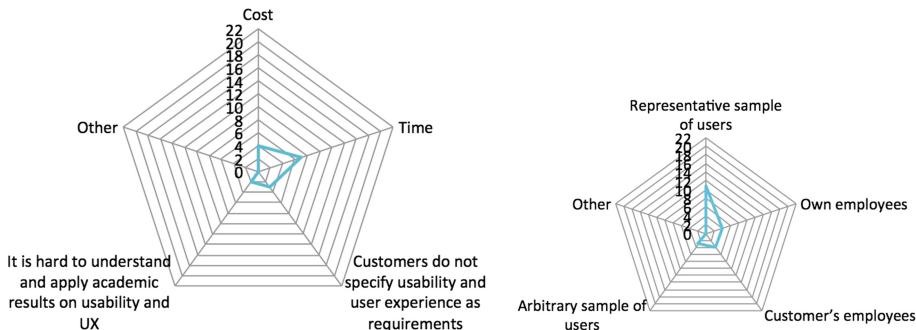
The results in Fig. 3 suggest that the organizations might not fully understand the distinction between user experience and usability issues. The results show that pragmatic aspects of user experience are prioritized by the organisations ahead of hedonic aspects such as fun/joy and helpfulness. These findings are also similar to the findings by [19].



**Fig. 3.** How UX is perceived in the respondents' organizations (multiple options allowed)

When asked about the frequency of conducting usability testing in projects, 12 organisations (55 %) indicate they always conduct usability testing and 10 organisations (45 %), indicate they sometimes do. Figure 4 (left-hand side) show some reasons why usability testing is not always conducted in the ten organisations.

The major reason indicated by the ten organisations was time constraints. This is consistent by the findings of [3]. However, software organisations that neglect usability



**Fig. 4.** Reasons for not conducting usability tests and the types of users selected for usability tests when they were conducted

aspects due to time or cost constraints, often spend more time and money on training and fixing bugs [14, 16].

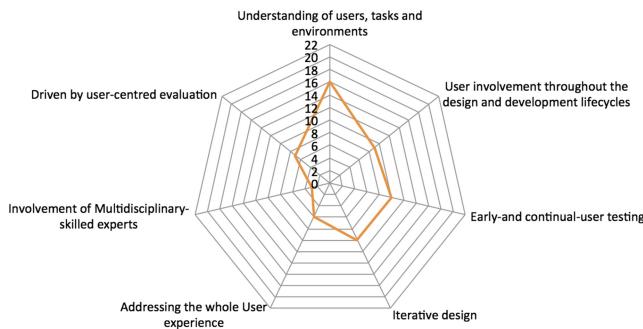
Respondents also provide information regarding the kind of users selected for usability testing in their organizations. This is shown in Fig. 4 on the right. As it can be seen, eleven organisations (50 %) indicate they use a representative sample of users for usability testing. Four organisations (18 %), used their own employees, four organisations (18 %), used their customer's employees and three organisations (14 %), used an arbitrary sample of actual users.

Similarly, some challenges as presented in Table 1 suggest reasons for the choice of RAD in most of the organizations. Rapid Application Development is particularly used where organisations are constrained by a short time frame for software projects, cost, and need for quality [1]. However, organizations suffer when they substitute HCI aspects for time and cost [7]. A respondent reports: “*It's time consuming.*”

**Table 1.** Challenges indicated for conducting usability, user experience and HCD Practices

Challenges	Respondents	Percentage (%)
Lack of standard tools for integration	2	9 %
Lack of knowledge of best practices	5	23 %
Short time to deploy software projects	7	32 %
Cost of hiring HCI experts	7	32 %
Ineffective government policies	1	4 %

We requested the organisations to select the principles for HCD, being applied by them, according to the ISO 9241-210 framework [11]. The results obtained and presented in Fig. 5 show that the only aspect that is more prioritized is the understanding of users, tasks and environments. Other important dimensions such as user involvement and involvement of multidisciplinary-skilled experts are less prioritized. Our results in this regard, are similar to the study by [13].



**Fig. 5.** HCD principles used in respondents' organizations (multiple options allowed)

### 3.3 Interviews

We conducted ten semi-structured interviews with software practitioners in three companies. Two of the companies are small-sized and are involved with custom development. The other company is medium-sized and is involved with off-the-shelf development. We wanted to know about interviewees' educational background. None of the interviewees, including those saddled with user interface design, have strong HCI backgrounds. At their best, these interviewees have developed their knowledge of HCI from reading HCI books and web descriptions of HCI practices such as usability. One of the interviewees, a Chief Technical Officer (CTO) in one of the small-sized companies, responded: "*In trying to protect ourselves, we found ourselves implementing HCI.*" The insight we can draw here is that of an educational limit.

Regarding user experience and human-centred design practices, we can say that the practitioners' organizations are not implementing these HCI practices, as we would expect. There are a couple of issues to justify our assumption. For example, none of the companies have a UX expert and no interaction design labs exist. The Project Manager (PM) in the medium-sized company attests: "*one area that we know we need to still work on is even our own development standards; we have not accepted that we are there, (in terms of) our benchmarks...*" The PM also indicated their challenge in doing UX work: "*we don't have a lab to say 'change the colour and see what people react to?' we don't do scientific experiments to decide what is the best colour scheme for this and that.*"

Regarding usability practices, again, we see some limitations in all of the three companies. As explained by the CTO of the medium-sized company, although the company uses focus groups for product conception and evaluation, there is bias in the user selection. The CTO hints: "*we get people within the company at different levels of IT to come and do like a focus group and run through this piece of software and throw up their challenges.*" We do not agree that that using our own employees would provide objective assessments for obvious reasons. In one of the small companies, a programmer, when asked about when to involve end-users in usability testing, responded: "*when the software is about to be implemented.*" The PM in the medium-sized company corroborates by indicating that: "*end-users are coming later into things that have been developed in many cases.*"

We wanted to know more about the importance of end-user involvement in software projects. The PM in the medium-sized company gave a hint: “*Where I see the involvement of the users would be because in a project for example, we are dealing with a specified level of users, maybe the senior users, the technical users, within the company to accept your product.*” We therefore, asked the company’s perception of the end-user. The response is: “*Those are the people (the senior users and technical users) I call end-users per se.*”

Finally, we found out through the interviews, that none of the three companies was familiar with the ISO 9241-210 framework for human-centred design. Thus, the companies only used their own methods. However, we found promise for the adoption of HCI practices in these companies. All the companies are aware of HCI, strive to build intuitive and visually appealing products, albeit the companies lack the expertise to engage in more productive HCI practices.

## 4 Discussions

The overarching insight drawn from our study is that there is a major gap between what HCI practice is elsewhere and how it is currently practiced in the Nigerian software companies investigated. It appears that HCI education and practice in Nigeria so far, are at the same level. Although HCI awareness is there and there is basic knowledge of HCI being applied in the industry, several issues exist which limit the uptake of HCI practices in Nigeria.

End users’ involvement is lacking in the way software development is carried out in all the companies investigated. As Nigeria has embraced a cashless economy, people involvement in software projects is critical. A major problem regarding end user involvement in Nigeria could be that of the perception of who the end user actually is. So far, we have not seen it documented in the literature that the end-users should be technical people and senior employees. The primary purpose for users involvement seems to be endorsement of a product. This would not be ideal for an information society.

Government policies are reported to be plodding and ineffectual. For example, the current software policy is yet to be enforced and many of the government actions are spontaneous. However, as can be seen in the introduction, the software industry in Nigeria is still in its formative stage and it is clear that standards and regulations are lacking as well.

The level of HCI knowledge in Nigeria currently, is limited. Thus, the industry is not able to give enough to meet the demands from the market. There could be a need to review the HCI education curriculum to ensure that it matches global standards. Unlike in the developed countries where HCI practices have advanced, the story is quite different in developing regions such as Africa (see e.g. [9]). In a recent study conducted in Colombia by Collazos and Merchan, [4], the need was stressed for local universities to “develop real-world projects as experimental studies that consider industry needs, bringing together participants from both academia and industry” [p.8]. Similarly, Winschiers, [20], in a study in Namibia emphasized that “methods have to be evaluated within the design process and adopted to the context” [p.75].

However, in comparison to other developing countries such as India and China where HCI practice is growing rapidly [17], the same cannot be said of Nigeria. At best, HCI is just at the awareness level in Nigerian software companies. However, the Nigerian market environment seems to be driving these software companies towards fully taking up HCI practices. A CTO of one of the small companies, when asked how they came to be aware of HCI, responded: “*maybe the market taught us.*”

This study has a major limitation. Although we strove to get many participants for the survey, our sample is not representative. This is still a major challenge to most quantitative studies [5].

In future work, we plan to investigate HCI education in Nigerian universities.

**Acknowledgments.** This research was supported by European Social Fund’s Doctoral Studies and Internationalisation Programme DoRa, which is carried out by Foundation Archimedes.

## References

1. Agarwal, R., Prasad, J., Tanniru, M., Lynch, J.: Risks of rapid application development. *Commun. ACM* **43**(11), 177–188 (2000)
2. Akintola, K.G., Akinyede, R.O., Agbonifo, C.O.: Appraising Nigeria readiness for e-commerce towards achieving vision 20:20. *Int. J. Res. Rev. Appl. Sci.* **9**(2), 330–340 (2011)
3. Arditò, C., Buono, P., Caivano, D., Costabile, M.F., Lanzilotti, R.: Investigating and promoting UX practice in industry: an experimental study. *Int. J. Hum. Comput. Stud.* **72**(6), 542–551 (2014)
4. Collazos, C.A., Merchan, L.: Human-computer interaction in Colombia: bridging the gap between education and industry. *IT Emerg. Mark.* **38**(6), 900–915 (2013)
5. Cycyota, C.S., Harrison, D.A.: What (not) to expect when surveying executives: a meta-analysis of top manager response rates and techniques over time. *Organ. Res. Methods.* **9**(2), 133–160 (2006)
6. Egbokhare, F.A.: Causes of software/information technology project failures in nigerian software development organizations. *African J. Comput. ICT* **7**(2), 107–110 (2014)
7. Gulliksen, J., Göransson, B., Bovie, I.: Key principles for user-centred systems design. *Behav. Inf. Technol.* **22**(6), 397–409 (2003)
8. Hotez, P.J., Asojo, O., Adesina, M.: Nigeria: “ground zero” for the high prevalence neglected tropical diseases. *PLoS Negl. Trop. Dis.* **6**, 7 (2012)
9. Hussain, Z., Slany, W., Holzinger, A.: Current state of agile user-centered design: a survey. In: Holzinger, A., Miesenberger, K. (eds.) *USAB 2009. LNCS*, vol. 5889, pp. 416–427. Springer, Heidelberg (2009)
10. Hussey, J., Hussey, R.: *Business Research: A Practical Guide for Undergraduate and Postgraduate Students*. Macmillan, London (1997)
11. ISO: Ergonomics of Human-System Interaction - Part 210: Human-Centred Design for Interactive Systems. ISO 9241-210:2010, pp. 1–32. ISO (2010)
12. Ji, Y.G., Yun, M.H.: Enhancing the minority discipline in the it industry: a survey of usability and user-centered design practice. *Int. J. Human- Comput. Interact.* **20**(2), 117–134 (2006)
13. Larusdottir, M.K., Haraldsdottir, O., Mikkelsen, B.: User involvement in icelandic software industry. In: *Proceedings of the INTERACT 2009*, pp. 1–2. ACM, Uppsala (2009)

14. Lizano, F., Sandoval, M.M., Bruun, A., Stage, J.: Usability evaluation in a digitally emerging country: a survey study. In: Proceedings of the INTERACT 2013, pp. 298–305. Springer, Cape Town (2013)
15. Phillips, C., Kemp, E.: The integration of HCI and software engineering. In: Proceedings of the ICSE 1998: (Education and Practice), pp. 399–401. IEEE Comput. Soc (1998)
16. Shneiderman, B., Plaisant, C.: Designing the User Interface: Strategies for Effective Human-Computer Interaction. Pearson Education Inc., New Jersey (2005)
17. Smith, A., Joshi, A., Liu, Z., Banon, L., Gulliksen, J., Li, C.: Institutionalizing HCI in Asia. In: Proceedings of the INTERACT 2007, pp. 85–99. Springer, Rio de Janeiro (2007)
18. Soriyan, H.A., Heeks, R.: A Profile of Nigeria's Software Industry. Precinct Centre, Manchester (2004)
19. Wechsung, I., Naumann, R., Schleicher, R.: Views on usability and user experience: from theory and practice. In: Proceedings of the NordiCHI 2008 Conference, pp. 1–4. ACM, Lund (2008)
20. Winschiers, H.: The challenges of participatory design in an intercultural context: designing for usability in Namibia. In: Proceedings of the Participatory Design Conference, Trento, Italy, Vol. II, pp. 73–76 (2006)

# Penan's Oroo' Short Message Signs (PO-SMS): Co-design of a Digital Jungle Sign Language Application

Tariq Zaman<sup>1</sup>(✉) and Heike Winschiers-Theophilus<sup>2</sup>

<sup>1</sup> Institute of Social Informatics and Technological Innovations,  
Universiti Malaysia Sarawak, Kota Samarahan, Malaysia

[zamantariq@gmail.com](mailto:zamantariq@gmail.com)

<sup>2</sup> Polytechnic of Namibia, Windhoek, Namibia  
[hwinschiers@polytechnic.edu.na](mailto:hwinschiers@polytechnic.edu.na)

**Abstract.** Oroo', a very peculiar jungle sign language of the semi-nomadic Penan in Malaysia, Borneo Island, is at the virtue of extinction with recent changes in lifestyle. The youth inhabiting the rainforest are more drawn to technology than traditional forest activities needing cognizance of Oroo'. In partnership with community members of Long Lamai, as part of a long term collaboration, we launched into revitalizing Oroo' through digitalization. Complementing previous efforts of database, tangible and game developments, we postulate that a language can only be revitalized if integrated in daily use. Thus in this paper we present the co-design of the Penan's Oroo' Short Message Signs (PO-SMS) application, which extends current technology driven communication means. Following a community-based co-design approach, a group of local youth and elders have led the unique design of their own digital indigenous communication tool. Our research contributes directly to the INTERACT 2015 theme of "Connection.Tradition.Innovation".

**Keywords:** Indigenous language · Community-based Co-design · Oroo' · Local content creation

## 1 Introduction

*"Many linguists predict that at least half of the world's 6,000 or so languages will be dead or dying by the year 2050. Languages are becoming extinct at twice the rate of endangered mammals and four times the rate of endangered birds"* ([1] as cited in [2]).

Countless indigenous languages have emerged over epochs of which many have already disappeared and others are at the virtue of extinction. Besides spoken and written languages, a great variation of alternative local language systems such as sign languages have evolved. The languages vary by degrees of complexity and strengths of expression. The most developed and current sign languages are without doubt the deaf signing languages which are composed of gestures symbolizing words or letters. Another well-known sign communication is composed of smoke signals, one of the oldest forms of long-distance visual communication. The Native Americans have developed complex

signals with dependencies of locations from where it is sent to ensure their enemies cannot decipher the messages. While for example the Vatican's papal election procedure uses a simple binary smoke code to announce a positive or negative result. Other unique languages are whistled and whistling languages, emulating a tonal oral language, such as Silbo Gomero practiced in Spain, and other variations found in Mexico and some African countries. However, even less noticed are the rainforest sign languages, which have been used by different tribes on Borneo Island. The language is composed out of different natural elements found in the jungle and combined along so called message sticks. Numerous different messages can be formed through combinations of signs, expressing warnings, information of whereabouts, state and activities [3, 4]. However less and less inhabitants of the Borneo rainforest master these languages due to changes in life conditions. Because most tribes have meanwhile settled and are no longer nomadic, the need for excessive messaging in the forest has diminished. The younger generation is no longer interested in upholding the knowledge of the language but rather drawn to technology and other modern utilities.

Languages are major carriers of culture and worldviews, which must be preserved primarily from a cultural heritage preservation perspective. *"In contrast to dominant languages with long written traditions, most indigenous languages either lost their written systems during colonization or never had one. The current struggle involves standardizing a written system, and the challenges are linguistic, political, and cultural"* [2]. Thus most early attempts to digitalize indigenous languages were dominated by database and website approaches, which were based on transcribing local languages. The major aim was to document and gather information around the languages, in projects such as the Australian Indigenous Language Database,<sup>1</sup> Indigenous Languages of the Americas library,<sup>2</sup> and the Archive of the Indigenous Languages of Latin America.<sup>3</sup> Yet their organization is neither inviting for native speakers, nor captures non-written languages nor encourages the active use of the language.

*"The definition of a healthy language is one that acquires new speakers. No matter how many adults use the language, if it isn't passed to the next generation, its fate is already sealed. Although a language may continue to exist for a long time as a second or ceremonial language, it is moribund as soon as children stop learning it"* ([1] as cited in [2]).

We postulate that languages can be sustained only if used in every day's life. With new trends and interaction styles of technologies, new opportunities have opened up to revitalize indigenous languages. Besides large audio and video repositories, a number of mobile apps and games have been developed. Yet connecting indigenous people, via their own traditional forms of communication has not been prominent in current trends of innovation and technology design.

In this paper we present one of our recent technology co-design ventures, namely a jungle sign language app development, which we have conceptualized with a semi-nomadic Penan community in Long Lamai, Malaysia, Borneo Island. The endeavor is part of a larger project on indigenous knowledge management and language

---

<sup>1</sup> (<http://austlang.aiatsis.gov.au/main.php>).

<sup>2</sup> ([http://www.brown.edu/Facilities/John\\_Carter\\_Brown\\_Library/ildb/index.php](http://www.brown.edu/Facilities/John_Carter_Brown_Library/ildb/index.php)).

<sup>3</sup> (<http://www.lib.utexas.edu/indexes/titles.php?id=16>).

preservation in a long term community development collaboration established between the Institute of Social Informatics and Technological Innovations (ISITI) at the Universiti Malaysia Sarawak (UNIMAS), and the Long Lamai community. In this paper we first explore current digitalization efforts of indigenous languages especially sign languages. We then provide the research and project context as well as our documentation and digitalization efforts of Oroo', the sign language of the Penan. We then describe the co-design process and outcome of the Penan's Oroo' SMS app. We conclude with a short reflection on our co-design and contribution to the field of local content creation under the theme of "Connection.Tradition.Innovation".

## 2 Revitalizing Indigenous (Sign) Languages

### 2.1 Digitalizing Indigenous Languages

Many initiatives of culture and language preservation have mushroomed over the last decade. Most have followed a traditional technical solution in form of a databases or a website, with the purpose of documentation and cultural heritage preservation rather than promoting the usage of the language. Often the development and even the use of the technology were detached from the native speakers.

*"To get ahead in the modern world without losing their heritage, indigenous communities need to develop a biculturalism that enables them to move between two cultures and to combine certain elements of each harmoniously. [...] In a digital world, it means Internet chats in indigenous languages, indigenous web pages, multimedia CD-ROMs for learning indigenous languages, and cultural information published by indigenous groups for a global audience" [2].*

Thus recently more initiatives aimed at developing technologies for or even with indigenous communities. A good demonstration of such an endeavor is the Ara Irititja Knowledge Management System (<http://www.irititja.com/index.html>) which brings back materials of cultural and historical significance to Anangu speaking people in Central Australia. Similar commendable are projects such as "The Yanomami Intercultural Education Program" where indigenous communities were involved in the development of digital educational material thereby enhancing the literacy rate in indigenous languages [5]. While the examples are many relatively few indigenous "non-written" languages are digitalized.

### 2.2 Digitalization Efforts of Indigenous Sign Languages

Sand drawings are a medium of communication among the members of the various language groups living in the north of the Vanuatu archipelago in the South Pacific Ocean. Sand drawings uniquely express the deep understanding of the land, conveying a sense of community, identity, and interaction with nature and history [6]. In 2011, in partnership with the French Embassy, the Vanuatu Cultural Centre signed an agreement for the digitalization of the audiovisual, photographic and sound archives of the Archive Unit. Therefore, more than 3000 audiotapes, 1000 VHS and others video format and about 3000 photography are currently digitalized. However the focus of the

digitalization process is mainly preservation of this wonderful cultural wealth rather than revitalization. The later could be supported with current touch-based drawing apps (such as “Sand Draw) with haptic feedback thereby simulating the drawing in the sand feeling.

An innovative use of newer technologies to rekindle smoke signals as they were used by Native Americans has been developed by Dennis Be Bel. Native Americans used smoke signals from smudge fires to exchange information over great distances and across cultures [7]. Dennis Be Bel developed smoke messaging service iPhone add-on hardware case that will release puffs of smoke to communicate securely with others [8]. At the push of a button, lamp-oil is heated and vaporized, sending a little cloud of smoke up in the air. This iPhone cover allows two people to speak to each other using agreed upon common and programmed code. So the users can assign their own code, number of puffs for a sentence such as two puffs for “how are you?” and three puffs for “I am fine”, however it will be difficult to manage complex conversation with this tool. The app in its current form rather has entertainment value than a serious language authenticity.

Another example is Silbo Gomero, the whistle language of La Gomera in Spain Canary Islands, off the coast of Morocco. Although whistled languages can be found around the world, they are rare, and few are likely to survive in the long term [9]. Silbo Gomero is not endangered but a unique indigenous way of communication. There are approximate 22,000 numbers of speakers or rather whistlers [10]. Silbo Gomero uses whistles meant to mimic the sounds of four vowels and four consonants, which, when used in conjunction, are able to create a vocabulary of over 4,000 intelligible words. Silbo Gomero can be understood at a distance of up to two miles, much further and with less effort than shouting. Some of the highly skilled whistlers are able to send messages from one end of the island to another [11]. Complementing government efforts of teaching the whistling language, Busuu is a social network for learning languages and based on a freemium business model. The website (with downloadable app) provides 80 learning units for thirteen languages including Silbo Gomero in Spanish, French, German and English.

### 3 Project Context

Digitalizing Oroo’ is a collaborative project of ISITI and Long Lamai, a local Penan community in Malaysia, on Borneo Island. Penans are one of the indigenous communities living in Sarawak, Brunei and Kalimantan [12]. The Sarawak Penan population in 2010 was estimated to be 16,281 people of whom about 77 % have settled permanently. The remaining 20 % are semi-nomadic while 3 % are still nomadic [13]. Long Lamai is one of the most progressive Penan communities in the upper reaches of Sarawak’s Baram river basin. It is very remote, requiring one and a half hours’ flight from Miri, and then an hours’ longboat journey upriver. There are 105 households and a population of app. 500. There is no 24-hour electricity supply and limited telecommunication service. Some families have generator sets to generate power, but few families can afford this. The Penan in Long Lamai were nomads, but have settled down in the area for over 50 years. Today the communities in Long Lamai is mainly involved in subsistence farming, and thus face issues of urban migration of their youths, reduced

opportunities to economic activities to improve livelihoods and loss of their indigenous knowledge. In this light the ISITI has engaged in a number of projects with the Long Lamai community, such as the e-Lamai Telecentre which was inaugurated in 2009, supporting initiatives such as the development of an indigenous botanical knowledge repository, an online Penan language dictionary and an e-health system. A number of other joint cross-disciplinary research and outreach activities are undertaken in the field of rural-based tourism, health and infrastructure in order to nurture a sustainable socio-economy.

In 2013, the ISITI and Polytechnic of Namibia with the collaboration of Long Lamai community initiated a project for digitalizing and preserving of Oroo'. Oroo' is a living cultural heritage of the Penan which from a historical, political, social and scientific perspective is of extreme value to society. The Oroo' project's main goal is to preserve the traditional knowledge and revive the sign language of the community, given that the older generation is slowly dying out, and knowledge is no longer being transferred to the younger generation. The younger generation although living in the remote village does not travel the forest any longer. Thus they are not interested in learning and retaining Oroo', as being the old way of communication. Over the last year we have assessed the number of signs known across the different age groups of the community. The elders know on average about 30 signs, while people between 30 to 41 years 10 signs, between 20 and 30 years 6 signs and the under the age of 20 years know about only 3 Oroo' signs [14]. Thus the digitalization efforts focus on documenting Oroo' signs and messages, creating a database (photos, video description, and drawings) and a rule system with the knowledgeable elders, while developing cultural Oroo' educational games for Penan children and ICT tools for contemporary use of Oroo' signs as communication medium for the youth.

### 3.1 Oroo' Documentation

**Rainforest Sign Languages in the Literature.** The rainforest sign languages of Borneo Island as documented and published by only a small number of anthropologists differ between the tribes yet have a common pattern. The jungle sign languages consist of a stick of varying length. Clefts cut into it hold a number of folded leaves, twigs and branches, which carry different meanings constituting the message [3, 4]. Messages such as event announcements, warnings, instructions and information can be communicated to other nomadic families of the same tribe (able to read the message). In case of the Murut's sign, the message contains the identity of the writer and placed in public spaces as public message [15]. These signs generally refer to ceremonial or hunting practices and describe the details of activities such as the direction of hunting, type, gender and size of hunted animal, weapon used for hunting, age, gender and personal or family affiliation of the writer. Few signs are for taboo in force, dangerous traps, marriage or death ceremonies, types of food given and the number of people attended the ceremony. Burrough reported that the Dusun (Kadazan) community uses signs and message sticks only to communicate with the spirit of their wet-paddi [3]. The Penan are recognized as the prodigies of jungle travel and sign reading [16]. The

Penan often traveled in groups, where the leading group left messages for the following once. Arnold [16] illustrates a few messages, demonstrating the richness of expression, e.g. messages such as: *The first group waited a long time for the second concerning an urgent matter. Thus the second group is now requested to travel through the night to catch up.*

In general, the literature references to the rainforest sign language hardly exceeded one paragraph and a number of examples of messages illustrated by drawings.

**Sign Collection.** Considering the extremely scarce and incomprehensive Oroo' documentation in the literature, we have launched into our own collection of signs, messages and rules. We have over the last two years collected about 50 different signs during a number of jungle walks around Long Lamai. Community members demonstrated the different Oroo' signs and messages (Fig. 1). We have video recorded the process of making the signs and taken photos of each one. We have documented the name and description of each sign and combinations thereof. At each visit an elder constructed Oroo' signs and messages as they came to his mind in the real environment. The authenticity was verified by a second accompanying elder. Considering the density of the jungle, the clarity and details of the signs and messages are at times blurred on the digital photos, as other twigs, branches and leaves are all around. Thus, a local artists group was engaged to draw each Oroo' sign. The drawings helped to understand the detailed complexities of the Oroo' signs (Figs. 2 and 3). The drawings are used alongside the photographs for further discussions, documentations and categorizations. All the drawings, photos and descriptions were reconfirmed by the community elders in community meetings.

**Sign Combinations.** In the absence of a documented grammar we have recently engaged in a systematic exercise to explicate the tacit rules of combining the signs forming a valid message, as well as categorizing the signs. While at this point our analysis is still incomplete, we have uncovered a number of subtle variations and meanings in the message composition. Firstly the 'Batang Oroo' (or 'message stick') which shows the direction where the sign maker (writer) is going and also has different positions where signs can be attached and thereby combined to a full message (see Fig. 3).

We were shown only three signs that are placed on the ground next to 'Batang Oroo' set obliquely in the ground; namely 'Pelun' the accumulation of leaves with the meaning of "wait here for me/us" and 'Tebai', a v-stick pointing at a direction saying



**Fig. 1.** Elders creating Oroo' messages



**Fig. 2.** Drawn Oroo' sign for "hunted wild boar"



**Fig. 3.** I'm alone, very very hungry and I have only water to survive. We are friends. I am going to the river



**Fig. 4.** Elders grouping Ooro' sign cards

“follow me this way” and ‘Selikang’, two straight sticks crossed at the midpoint and one message stick saying “don’t go in this direction”. All three being direct instructions to the reader of the message. Then a number of clefts along the stick with one cleft at knee level, mostly used to indicate number of people, state (hungry, thirsty). Other clefts app.  $\frac{3}{4}$  into the message stick and a cleft at the end of the stick are used for activities, durations and destinations. The full description of the rule system is beyond the scope of this paper.

**Sign Grouping.** In order to understand the underlying categorization of the signs we have engaged a group of 5 elders in the jungle in a card sorting activity. [17, 18] have demonstrated that card sorting in a cross-cultural design context can reveal cultural adequate mental models of categorization necessary for local adaptions of design. Following a generative approach all signs (either drawn or photographed, depending on the quality) were printed on individual cards and displayed all at the same time to the elders (Fig. 4). They were requested to group what “belongs together” according to their perception. The elders immediately understood the exercise and launched into piling up related concepts. A few discussions with a few re-groupings led to a preliminary set of 16 named categories. Asked to reduce the number of categories, the elders put together some of the piles and re-allocated a name to the new categories. A total of 9 categories were formed, namely Instructions, Warnings, Information, Directions, Durations, Animals (animals, hunting), Number of people, River place activity (river, fishing), Hungry. The categories have not yet been verified at a community meeting.

**New Sign Creation.** Another interesting aspect was in the exploration of existing signs and “missing” signs. While there are signs for death announcement, differentiated by sex and age there are no signs for birth announcements nor for the distinction of sex for life persons. However, within the discussion the elders with great ease expanded the current existing set of signs by “birth of baby boy/girl” and woman and man. Following the logic of other signs, consensus on the signs was immediate. Thus the woman was symbolized by the sign of river as she is the one fishing, while the man was symbolized

by a blow pipe as he is the one hunting, following the same association of the representation of dead woman being an ‘Atip lutan’ (fire tong, as women are the ones who make fire) and dead man being an ‘Atip na’o’ (utensil used by man for eating starchy sago flour).

### 3.2 Previous Digitalization Efforts

Besides having created a database of current signs and messages composed of the Penan term, English translation, photo, drawing and video recording of construction and explanations, we have evaluated so far two different approaches to enhancing the learning of signs with technology.

**Tangibles.** Capacitive sensing tangibles linked with 2D representations on tablets, were explored as a novel concept for teaching Ooro’ signs. The tangibles were constructed out of a baseplate with touch points topped with the Ooro’ sign made out of clay and natural products. Their sole function programmed was to initialize the signs on a tablet for further manipulation of the 2-dimensional digital representations of complex messages. The tangible-tablet tool was evaluated as a collaborative learning tool in the village where 6 family groups, consisting of respectively an Ooro’ conversant parent and a child engaged with the tool [19].

Although the tool seemed engaging and the children learned most of the 10 signs tested for, a number of critical points show the unwieldiness of the tool. Firstly tangibles constructed were only functional for a few days, due to weather conditions. Using fresh leaves is impractical as they die after a few days, which modifies the authenticity of the sign; such as the ‘wild boar’ can only be symbolized by a fresh leaf. The creation of permanent artificial signs would resolve this, yet in general the question regarding the purpose or use and quality of these handmade tangibles using capacitive sensing remains. The technology chosen does not allow for any further interaction with the signs then the retrieval of the digital sign, which substantially undermines the characteristic of the language. There is no evidence that beyond the novelty effect the tangibles would actually enhance learning. Through discussions with community elders, we realized the complexities of interaction between Ooro’ signs, tangible tools and the corresponding two dimensional images.

**Games.** We have developed an Ooro’ adventure PC game with three stages for kids. On the first level the Ooro’ signs are explained in the form of stories followed by an interactive component of finding hidden signs scattered in the background picture of the rainforest. Then, the users have to distinguish between pairs of presented Ooro’ messages. And finally the kids are presented with an Ooro’ sign representing an animal and are requested to shoot the corresponding animal with a blow pipe. The game is then followed by a quiz to test the users’ Ooro’ sign knowledge. A first evaluation, with 17 kids in Long Lamai, suggests that the general interest in local content games is awakened yet the learning curve has been rather shallow. A number of improvements are required to enhance the performance and its applicability beyond kids [14].

## 4 PO-SMS

We postulate that a language is only truly sustained if in use. Thus we have conceptualized the digitalization of the Oroo' language, as an extension to current means of communications such as sms or whatsapp. The elders have expressed their concern about the lack of knowledge and lack of interest by the youth in learning the language yet being drawn towards technology-driven communications. Thus the idea of an Oroo' sms app arose within the discussions in the village. In a pre-study with a selected group of youth a number of new Oroo' signs were created for assumed messages to be conveyed to each other via cell phones, such as "How are you?", "I am well/not feeling well", "Where are you?", "I am at home/school/church", "I am taking dinner", "I am going for fishing", "I am cooking". The pre-study showed promising results and interest by the youth, thus the development was endorsed.

### 4.1 Community-Based Co-design Approach

To ensure a higher authenticity and user acceptance the development of the new app follows a community-based co-design approach [20]. The methodology is based on principles of participatory design with adapted methods to the community context. We would like to emphasize that the first author has been regularly returning to the community for the last years thereby establishing a strong trust relationship. He has worked on a number of projects with this community among others the establishment of cultural protocols and the preservation of indigenous knowledge both providing contributions to local interaction design methods [21, 22].

### 4.2 Co-design Workshop

In late 2014 a two-day workshop was held with eight youth in Long Lamai, three female and five male participants in the age group between 14 and 24, as well as one male participant of age 32. The workshop consisted of different phases. In the first phase personas were created, representing Long Lamai community members. Then paper SMS were created as sent between the personas to capture typical and realistic sms texts. The sms'es were categorized and a reasonable number were selected for which Oroo' signs were created (if necessary) following the logic of the existing signs. Then signs were categorized to define a user interface organization and the layout of the interface was determined jointly with a final paper prototype as outcome.

**Personas.** A relatively popular user-centered design method is the use of personas, which are describing a fictitious future user of a system to be developed. Benefits are seen as "*ranging from increasing the focus on users and their needs, to being an effective communication tool, to having direct design influence, such as leading to better design decisions and defining the product's feature*" [23]. Mostly personas are created by the designers as an amalgamation of collected data about users thereby

<b>Vasper</b> -26 years -Camera Man -From Long Lamai -So handsome -Like Jungle Trekking -Funny Guy -Friend of us -Love Long Lamai and Jungle -Skinny man -Hardworking man -Creative man -Christian man	<b>Rebecca</b> -18 years old -Volleyball -Singing -Friendly -Hillary's friend -Working at the Telecentre -From Long Lamai -Loves creating handicraft -Loves surfing internet	<b>Clarisse</b> -25 years old -Working at Bario -Love Long Lamai -From Long Lamai -So friendly -Very Kind -So pretty -Hard work -Know how to cook -Like fishing and singing -Friend with Erna, Shida, Azlyn
<b>Damon</b> -20 years old -Pastor -Football -Fishing -Reading -Friendly -Vincent Friend	<b>Cassandra</b> -40 years -Handicraft -Housewife -Kind person -Loves to swim -Strict mother -Hardworking lady -From Long Banga -Love Long Lamai -Backup singer in church	<b>Alavick</b> -42 years old -Boat driver -Tourist guide -Reading Bible -Kind and ugly man -Garren's friend -From Long Puak -Working at clinic Long Banga

Fig. 5. Personas created by the youth

creating a typical user. We have adapted the method to participatory design principles, where the user participants, the Long Lamai youth, generated their own personas. The group of youth was split in a group of girls and a group of boys each creating a male and a female youth persona. Then as a full group they jointly described an older male and female persona. Thus a total of 6 personas were generated (Fig. 5). The primary aim of the personas was for the youth to sufficiently identify with the one or the other persona for the next phase of the workshop and to be able to compose sms texts freely without them being attributed to any participant.

**SMS Composing.** In the sms composing phase of the workshop, the personas now familiar to every participant were placed in the middle. Every participant had a pile of A6 cards next to them and was asked to compose as many sms texts as possible with the following rules: Only one sms per card indicating clearly which persona wrote to which other persona. The card could then either be dropped in a box (for no one to see), or handed over to another participant with the request to reply or put out in the open for any participant to read and reply. This phase of the workshop was without doubt the liveliest activity, where all participants engaged in writing sms, reading others, laughing out loud and replying promptly (Fig. 6(a)). Within less than 15 min 89 sms texts were created, with one third in Malay and the rest in English. A first grouping of identical or similar



**Fig. 6.** (a) and (b) Left: SMS composing, Right: Ooro' sign creation

sms texts showed the highest occurrences of the following messages: "What are you doing?" (7), meeting up (7), "what do you mean? (4), where about question (4). Considering the large number of sms texts created the group agreed on 13 sms texts to create Ooro' signs for during the workshop (see an excerpt in Table 1).

**Creation of Ooro' Signs.** Firstly the group went through a slide show of the documented Ooro' signs, to ensure that every participant recalls the existing signs. Then a couple of participants went to the nearby woods and collected a set of items, such as twigs, branches, and leaves. Jointly the signs were created one by one with trials, suggestions and lively discussions until consensus was achieved (Fig. 6(b)). The final sign was photographed and drawn by the participants.

Table 1 presents an excerpt of signs created and reused.

**Ooro' Signs Categorization.** Firstly the existing and newly created Ooro' signs were drawn on one card each by the participants. Then the participants were asked to group the cards as they "belong together", similarly to the elders' card sorting (see section sign grouping). The youth struggled to come up with an appropriate categorization but after a long time finally agreed to the following:

- Activities: waiting, fishing, meeting, pick nick, going
- State: are you hungry? hungry, not hungry, are you fine?, fine, not fine
- Objects: house, plane
- Living beings: person, monkey, boar, friends.

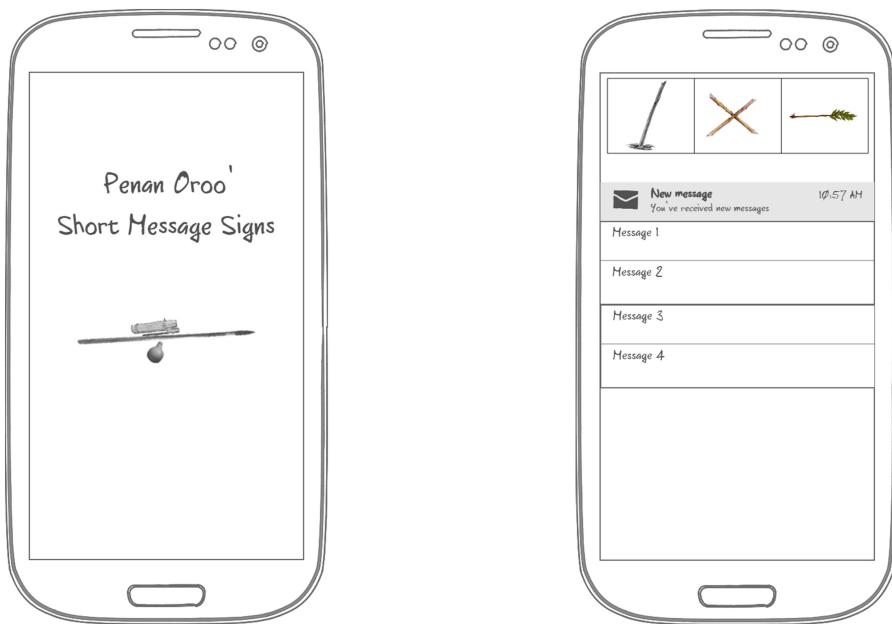
The categorization was tested on four community members that were not part of the grouping exercise. Each of them was shown the four categories and requested to point out in which category a specific Ooro' sign prompted would be. All community members identified the correct categories without hesitation. This 'quick and dirty' testing hints at a usable classification.

**App Screen Layout Design.** The layout is aiming for a middle sized touch screen smart phone considering that more than half of the participants own such a smartphone. The layout was co-sketched with the youth participants screen by screen exploring the sequence and interactions. The pictures below show the screen sequence and layout created.

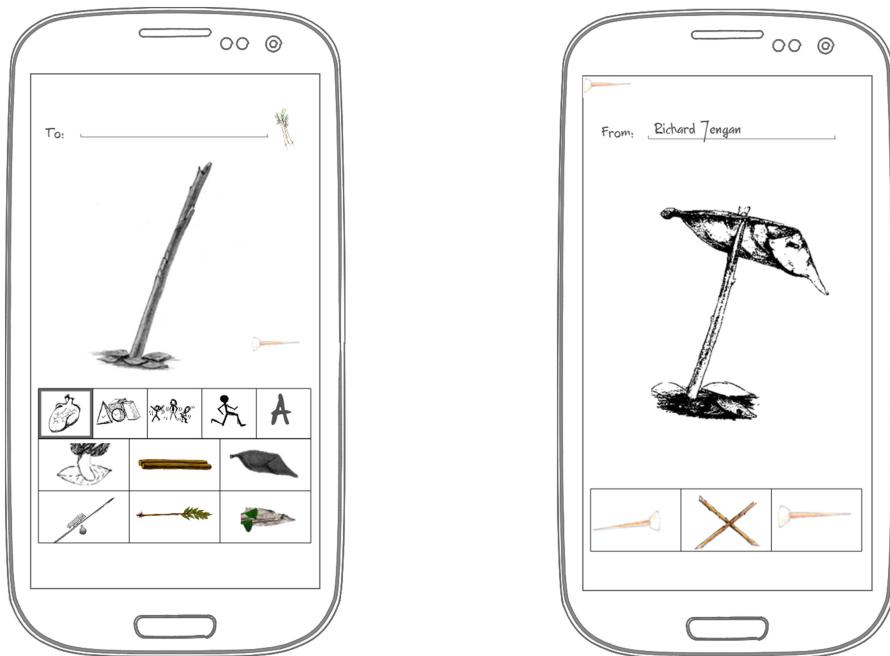
**Table 1.** Sample signs: newly created and existing ones

Number	SMS text	Sign	Comment
1	How are you?		Composed out of existing sign "not fine" combined with a new sign for "fine"
2	I am not fine		Existing sign for "sick" whereby the scratched area shows the seriousness of the illness.
3	Are you hungry?		Inspired by the existing "hungry" sign with the leave unfold halfway and inserted at the end of the stick rather than any cleft.
4	I am hungry		Existing sign of a folded leave signifying hunger.
5	I am not hungry		Existing sign of "hungry" combined with existing sign of "not"
6	Where are you going?		A stick with many branches indicating different directions
7	I am waiting for you at home		Composed out of existing waiting sign and existing "home" sign
8	I am sleeping		Newly created out of a new sign for "bed" (wooden bars) and an existing sign of "person"
9	Kuching		Cat symbol as the town Kuching is known for its association with cats
10	Pick Nick		New sign for "fire place" combined with existing sign of "boar"

- Screen 1: The “blow pipe with its container of darts” was chosen as the app splash screen icon, thereby representing that darts which are blown/send to a recipient and at the same time being peculiar to the Penans
- Screen 2: The “batang Oroo” was chosen to represent the “compose message button”. Once pressed screen 3 appears. The crossed sticks are to quit the process and the twig with roots is an existing Oroo’ sign for “Person” which leads to the address book. “New Message” indicates the unread message and “Message 2”, “Message 3” and “Message 4” indicate read messages
- Screen 3: The empty “batang Oroo” will be populated with the chosen signs. The four categories (Living being, Objects, Status and Activity) are buttons once pressed display the signs in that category. Alternatively the A button allows for additional text. The “right dart” can be used as send button
- Screen 4: “From” represents sender of the message. Once the message arrived and read the “right dart” button is pressed to forward message and “left dart” to reply message (Figs. 7 and 8)



**Fig. 7.** (a) and (b) Left: Splash screen of PO-SMS, Right: Inbox screen



**Fig. 8.** (a) and (b) Left: Compose message screen, Right: Read messages screen

## 5 Conclusion and Future Work

This paper presents the community-based co-design process of an innovative communication tool, as designed in collaboration with an indigenous community aiming at reviving a traditional sign language.

The process itself has shown to be very rewarding for community members and technologists, engaging the youth and the elders in a creative development. A separation of elders and youth was done intentional for the youth to feel free and not be dominated by the elders in their originality. The persona method showed to be extremely fruitful in the youth unleashing the SMS texting freely. Taking the youth through all the phases of design up to the screen layout ensured that they understood all the steps and are empowered to do design modifications at any level. For instance we are aware that the designed new signs still have to go through a number of community validations, as well as the ‘artificial’ categorization of the signs. Most problematic has shown to be the categorization of the signs, which is needed for a fast retrieval in the process of compiling a message. The different categorization approach between the elders and the youth requires us to further investigate a suitable mechanism during the app use. It is planned that at the next visit the prototype implementation will be evaluated in the community, revealing a number of usability and design modification requirements.

Overall we trust that the community initiated development of this tool will contribute to the revitalization of a forgotten jungle sign language. We are aware that the

current set of messages of Oroo' and SMS texts contains an extremely small common subset. Thus it is necessitating the creation of numerous new signs in order to cover current expressions. Nevertheless the designed communication tool will be re-connecting indigenous elders and youth via their own tradition in an innovative way.

**Acknowledgement.** We thank all the participants of the Long Lamai community for taking part in the project. This work was carried out with the aid of a grant from the Information Society Innovation Fund ISIF Asia (L18403/I03/00/PROJECT ORO) and Universiti Malaysia Sarawak under Postdoctoral Fellowship program (DPD/(I03)/2014(02).

## References

1. Ostler, R.: Disappearing language. *Futurist*. **33**(7), 16–21 (1999)
2. Lieberman, A.: Taking ownership: strengthening indigenous cultures and languages through the use of ICTs, LearnLink. <http://goo.gl/6333Am>
3. Burrough, P.A.: Stick signs in the sook plain. *Sabah Soc. J.* **V**(2), 83–97 (1970)
4. Burrough, P.A.: Message sticks used by Murut and Dusun people in Sabah. *J. Malays. Branch R. Asiat. Soc.* **48**(2), 119–123 (1975)
5. Gómez, G.G.: Computer technology and native literacy in the amazon rain forest. In: Dyson, L.E., Hendriks, M., Grant, S. (eds.) *Information Technology and Indigenous People*, pp. 117–119. IGI Global, Hershey (2007)
6. Zagala, S.: Vanuatu sand drawing. *Museum Int.* **56**(1–2), 32–35 (2004)
7. Smith, C., Ward, G.: *Indigenous Cultures in an Interconnected World*. UBC Press, Canada (2000)
8. SARC Communicator: Use your iPhone to send smoke signals. SARC Communicator, p. 1, Surrey Amateur Radio Club, Surrey (2014)
9. Vanderlip, C.: Silbo gomero and whistled languages. Grand Valley State University (2013). <http://goo.gl/fTAMis>
10. Ellison, K.: 8 Languages you've never heard of (and who actually speaks them) gadling (2015). <http://goo.gl/R5tJYg>
11. Plitt, L.: Silbo gomero: a whistling language revived. BBC NEWS (2013). <http://goo.gl/mNCXMS>
12. Sercombe, P.G.: Small worlds: the language ecology of the Penan in Borneo. In: Hornberger, N.H. (ed.) *Encyclopedia of Language and Education*, pp. 3068–3078. Springer, US (2008)
13. Lyndon, N., Er, A., Sivapalan, S., Ali, H., Rosniza, A., Azima, A., Junaidi, A., Fuad, M., Hussein, M.Y., Helmi, A.M.: The world-view of Penan community on quality of life. *Asian Soc. Sci.* **9**(14), 98–105 (2013)
14. Zaman, T., Winschiers-Theophilus, H., Yeo, A., Ting, L.C., Jengan, G.: Reviving an indigenous rainforest sign language: digital Oroo adventure game. In: International Conference on Information and Communication Technologies and Development, Singapore, May 2015
15. Polunin, I.: A note on visual non-literary methods of communication among the Muruts of NorthBorneo. *Man* **59**, 97–99 (1959)
16. Arnold, G.: Nomadic Penan of the upper Rejang (Plieran), Sarawak. *J. Malays. Branch R. Asiat. Soc.* **31**, 40–82 (1958)

17. Petrie, H., Power, C., Cairns, P., Seneler, C.: Using card sorts for understanding website information architectures: technological, methodological and cultural issues. In: Campos, P., Graham, N., Jorge, J., Nunes, N., Palanque, P., Winckler, M. (eds.) INTERACT 2011, Part IV. LNCS, vol. 6949, pp. 309–322. Springer, Heidelberg (2011)
18. Rodil, K., Rehm, M., Winschiers-Theophilus, H.: Homestead creator: using card sorting in search for culture-aware categorizations of interface objects. In: Winckler, M. (ed.) INTERACT 2013, Part I. LNCS, vol. 8117, pp. 437–444. Springer, Heidelberg (2013)
19. Plimmer, B., He, L., Zaman, T., Karunananayaka, K., Yeo, A.W., Jengan, G., Blagojevic, R., Yi-Luen, E.D.: New interaction tools for preserving an old language. In: 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI 2015), pp. 3493–3502. ACM, New York, NY, USA (2015)
20. Winschiers-Theophilus, H., Bidwell, N.J.: Toward an Afro-Centric indigenous HCI paradigm. *Int. J. Hum. Comput. Interact.* **29**(4), 243–255 (2013). Taylor & Francis UK
21. Zaman, T., Yeo, A.W.: Ensuring participatory design through free, prior and informed consent: a tale of indigenous knowledge management system. In: Saeed, S. (ed.) User-Centric Technology Design for Nonprofit and Civic Engagements, vol. 9, pp. 41–54. Springer International Publishing, Switzerland (2014)
22. Zaman, T., Yeo, A.W., Kulathuramaiyer, N.: Augmenting indigenous knowledge management with information and communication technology. *Int. J. Serv. Technol. Manag.* **19**(1/2/3), 137–148 (2013)
23. Nielsen, L.: Personas. In: Soegaard, M., Dam, R.F. (eds.) *The Encyclopedia of Human-Computer Interaction*, 2nd edn. The Interaction Design Foundation, Aarhus (2014)

# The Whodunit Challenge: Mobilizing the Crowd in India

Aditya Vashistha<sup>1(✉)</sup>, Rajan Vaish<sup>2</sup>, Edward Cutrell<sup>3</sup>,  
and William Thies<sup>3</sup>

<sup>1</sup> University of Washington, Seattle, USA  
[adityav@cs.washington.edu](mailto:adityav@cs.washington.edu)

<sup>2</sup> University of California, Santa Cruz, USA  
[rvaish@cs.ucsc.edu](mailto:rvaish@cs.ucsc.edu)

<sup>3</sup> Microsoft Research India, Bangalore, India  
[{cutrell, thies}@microsoft.com](mailto:{cutrell, thies}@microsoft.com)

**Abstract.** While there has been a surge of interest in mobilizing the crowd to solve large-scale time-critical challenges, to date such work has focused on high-income countries and Internet-based solutions. In developing countries, approaches for crowd mobilization are often broader and more diverse, utilizing not only the Internet but also face-to-face and mobile communications. In this paper, we describe the Whodunit Challenge, the first social mobilization contest to be launched in India. The contest enabled participation via basic mobile phones and required rapid formation of large teams in order to solve a fictional mystery case. The challenge encompassed 7,700 participants in a single day and was won by a university team in about 5 h. To understand teams' strategies and experiences, we conducted 84 phone interviews. While the Internet was an important tool for most teams, in contrast to prior challenges we also found heavy reliance on personal networks and offline communication channels. We synthesize these findings and offer recommendations for future crowd mobilization challenges targeting low-income environments in developing countries.

**Keywords:** Crowdsourcing · Crowd mobilization · HCI4D · ICT4D · India

## 1 Introduction

Recent years have witnessed the power of crowdsourcing as a tool for solving important societal challenges [1–4]. Of particular note are instances of crowd mobilization, where large groups of people work together in service of a common goal. A landmark demonstration of crowd mobilization is the DARPA Network Challenge, where teams competed to find 10 red balloons that were hidden across the United States [5]. The winning team found all the balloons in less than nine hours, utilizing a recursive incentive structure that rewarded participants both for joining the search as well as for growing the team [6]. Since then, mobilization exercises such as the Tag Challenge have shown that teams can locate people of interest across North America and Europe [7]. The MyHeartMap Challenge mapped over 1,500 defibrillators in Philadelphia County [8]. Authorities have also turned to crowd mobilization for help gathering intelligence

surrounding the London riots [9] and the Boston Marathon bombings [10], though the results have not been without pitfalls [11] and controversy [12].

One limitation of prior crowd mobilization studies is that they have focused exclusively on North America and Europe, where Internet penetration is so high that most teams pursue purely online strategies. However, in other areas of the world, the Internet remains only one of several complementary channels for effective mobilization of crowd. For example, in India, 1.2 % of households have broadband Internet access [13], but there are 929 million mobile subscribers, over 550 million viewers of television, and over 160 million listeners to radio [13, 14]. An SMS-based social network called SMS GupShup has 66 million subscribers in India [15]. Moreover, there is a rich oral tradition of conveying stories and information face-to-face. Environments such as the Indian railways – serving 175 million passengers every week [16] – provide fertile grounds for mobilizing crowds. India also has a unique social milieu, with its own social hierarchies, attitudes towards privacy [17], and trust in/responsiveness to various incentive schemes. In light of all these characteristics, it stands to reason that effective crowd mobilization in India would require broader and more inclusive techniques than in Western contexts.

To further explore the landscape of crowd mobilization in India, this paper reports on a new mobilization contest that was designed specifically for the Indian context. Dubbed the “Whodunit Challenge”, the contest enabled participation through mobile phones instead of via the Internet. The contest offered a Rs. 100,000 (USD 1,667) prize<sup>1</sup> for solving a fictional mystery case, in which teams were asked to gather five pieces of information: *Who*, *What*, *Where*, *When*, and *Why*. To participate, an individual had to send a missed call<sup>2</sup> to the contest phone number, which returned via SMS one of five *phrases*, each providing one of the pieces of information. Because some phrases were returned with low probability, and only one phrase was sent to each phone number irrespective of the number of missed calls received, participants needed to form teams of several hundred people in order to have a chance of winning.

The Whodunit Challenge attracted over 7,700 participants within the first day, and was won by a university team in just over five hours. To understand teams’ experiences and strategies, we conducted 84 phone interviews, covering most individuals who submitted 3 or more phrases or who received phrases sent with low probability. While many of the winning teams did utilize the Internet to mobilize the crowd for finding phrases, we also uncovered interesting cases that relied mainly on face-to-face or mobile communication. Unlike previous crowd mobilization challenges, many successful teams relied only on personal networks, rather than trying to incentivize strangers to help them search for phrases. Members of these teams were usually unaware of (or unmotivated by) the cash award.

In the remainder of this paper, we describe the design rationale, execution strategy, and detailed evaluation of the Whodunit Challenge. To the best of our knowledge, this is the first paper to describe a large-scale crowd mobilization contest in a developing-country

---

<sup>1</sup> In this paper, we use an exchange rate of 1 USD = Rs. 60.

<sup>2</sup> Sending a missed call refers to the practice of calling a number and hanging up before the recipient can answer [6].

context, exploring the portfolio of online and offline communication strategies that teams employed. We also offer recommendations to inform the design of future crowd mobilization challenges targeting low-income environments in developing countries.

## 2 Related Work

There is a vibrant conversation in the research community surrounding the future of crowd work [18]. Research that is most closely related to our work falls in two areas: crowd mobilization challenges and crowdsourcing in developing regions.

One of the most high-profile experiments in crowd mobilization was DARPA's Network Challenge, launched in 2009. By asking teams to find ten red balloons that were hidden across the United States, the challenge aimed to explore the power of the Internet and social networks in mobilizing large groups to solve difficult, time-critical problems [5]. The winning team, from MIT, located all of the balloons within nine hours [19] using a recursive incentive mechanism that rewarded people for reporting balloons and for recruiting others to look for balloons [6]. This approach was inspired by the work of Dodds et al. [20], which emphasizes the importance of individual financial incentives [21]. Cebrian and colleagues proved that MIT's incentive scheme is optimal in terms of minimizing the investment to recover information [22], and that it is robust to misinformation [23].

The DARPA Network Challenge seeded broad interest in the role of social networks in homeland security [24]. This led to a follow-up contest called the Tag Challenge from the U.S. Department of State [7], in which the task was to find five people across five cities and two continents within twelve hours [25]. The winning team found three of the five people and used an incentive scheme similar to the one that won the Network Challenge. Private firms and universities have also explored the potential of crowd mobilization. In 2009, Wired Magazine launched the Vanish Challenge [26] and in 2012, the University of Pennsylvania launched the MyHeartMap Challenge. The latter challenge saw over 300 participants who found and catalogued over 1,500 defibrillators in Philadelphia County [8]. However, to the best of our knowledge, there has not yet been any social mobilization contest with a focus on a developing country. There is a need to explore the landscape of crowd mobilization in developing countries and to identify the differences from crowd mobilization strategies observed in the developed world.

Researchers have also studied the potential and limitations of crowdsourcing in developing regions. Platforms such as txtEagle [27] and mClerk [28] aim to enable workers to earn supplemental income on low-end mobile phones. Others have examined the usage [29, 30] and non-usage [31] of Mechanical Turk in India, where approximately one third of Turkers reside. Efforts such as Ushahidi [32] and Mission 4636 in Haiti [33] have leveraged crowd workers to respond to crises in developing countries. Researchers have also explored the role of social networks such as Facebook [34] and SMS GupShup [35] in low-income environments.

### 3 The Whodunit Challenge

The Whodunit Challenge was an India-wide social mobilization contest that awarded 100,000 Rupees (USD 1,667) to the winner. The objective of the challenge was to understand mechanisms, incentives and mediums people in India use to mobilize large groups of people for a time-bounded task.

#### 3.1 Design Principles

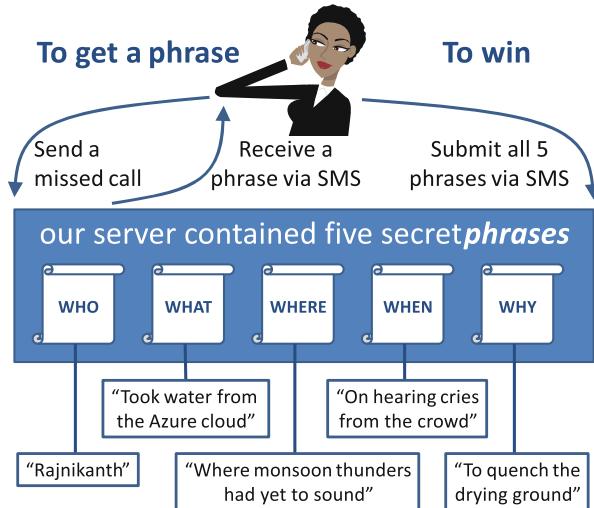
The Whodunit Challenge embodied three design principles to make it broadly accessible throughout India. In India, 72 % of the adult population is illiterate in English [36]. Thus, we localized the SMS messages by translating them into ten regional languages of India, making them more accessible than contests based on English alone. To ensure that the messages were not distorted in the translation, the translations were done by native speakers of local languages who were highly skilled in English. A majority of the Indian population has constrained access to modern devices and networks: the smartphone penetration is only 10 % [37] and Internet penetration is 20 % [38]. Thus, we aimed to enable participation by owners of basic mobile phones, thereby ruling out any dependence on computers, smart phones, or Internet connections (broadband or mobile). While Internet access could still offer advantages to participants, it was not strictly necessary to compete and win. Around 60 % of the Indian population earns less than US\$2 per day [39]. Thus, we aimed to minimize the costs of participation. To participate in the contest, users needed to send a missed call from a mobile phone (which incurs no cost to them). To submit a phrase, they needed to send an SMS; this costs at most US\$0.015, though is free under many mobile subscription plans. Our design did not require users to initiate any voice calls, as this expense could have thwarted participation from cost-sensitive groups.

#### 3.2 Contest Mechanics

The challenge required participants to reconstruct a secret sentence consisting of five pieces of information – Who, What, Where, When and Why (see Fig. 1). Each piece of information was referred to as a *phrase* and represented a part of the secret sentence.

To receive a phrase, participants simply sent a missed call to the contest phone number. On receiving the call, our server responded with an SMS containing one of the five phrases. Each phrase was sent in two languages: English and the predominant local language in the telecom circle from which the call was made. The first person to forward all five phrases (i.e., the secret sentence) to our server via SMS was declared the winner. User responses were passed through a transliteration API, providing robustness to any minor typos incurred in re-typing phrases.

What made the challenge difficult is that some phrases were very rare, thereby requiring participants to form large teams to gather all the phrases. Also, we made it difficult for any one person to receive many phrases by sending only a single phrase to each phone number even if we received multiple missed calls from the same number.



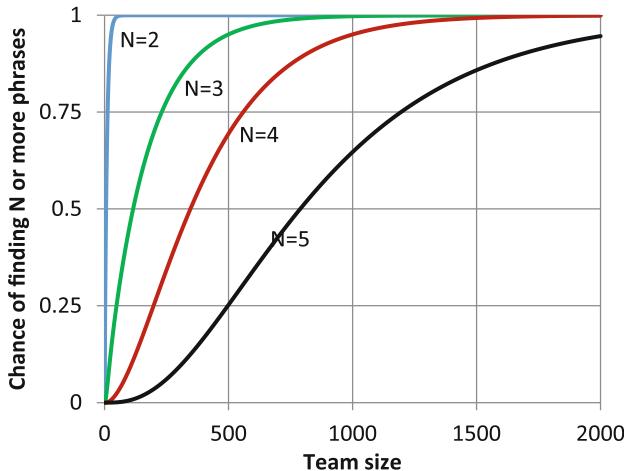
**Fig. 1.** Graphical illustration of the Whodunit Challenge

Regulations in India make it difficult for a person to obtain many phone numbers; for example, VoIP DID numbers are not available for sale (and our server ignored VoIP calls anyway). Also, telecom operators offer a limited number of SIMs per customer, and each requires several pages of paperwork and supporting documents (personal identification, proof of address, etc.). While we advised participants that a very large team would be necessary to win, the award itself was made to an individual. Thus, any sharing of the award within a team would need to be managed by a team leader.

While the Whodunit Challenge was framed in lighthearted terms, we intended for the search for phrases to closely mirror the search for serious time-sensitive information, such as missing persons, suspicious containers, counterfeit currencies, etc. By using electronic phrases instead of physical artifacts, we were able to monitor and control each step of the contest.

### 3.3 Chance of Winning

How large of a team was needed in order to win the challenge? We did not publicize this information broadly, though during one Q&A session, we indicated that competitive teams would contain several hundred members. In response to each missed call, the server responded according to a weighted random function, returning *Who*, *What*, *Where*, *When* and *Why* with probability 89.4 %, 10 %, 0.2 %, 0.2 %, and 0.2 %, respectively. Given these probabilities, the chance of winning as a function of team size is illustrated in Fig. 2. To have a 50 % chance of winning, a team needed 789 people. However, depending on their luck, smaller or larger teams could also win. To have a 5 % chance of winning, a team needed about 230 people; for a 95 % chance of winning, a team needed about 2040 people. The probability of winning did not depend



**Fig. 2.** Chance of finding  $N$  or more phrases as a function of team size

on participants' location, time of sending a missed call, or other factors, as each phrase was returned independently at random.

### 3.4 Publicity and Outreach

We publicized the challenge widely in order to seed participation. A distinguished speaker announced the challenge to a live audience of 2,500 undergraduate engineering students about one week prior to the contest launch [40]. We conducted a large email and social media campaign targeting engineering colleges, MBA colleges, and student volunteers connected with Microsoft Research India. We also presented posters at two academic conferences in the month preceding the contest to create awareness among computer scientists. While the audiences for these activities were primarily composed of Internet users, we advised team leaders that outreach to non-Internet users would be highly advantageous for growing a large team and winning the challenge. Also, to seed visibility among non-Internet users, we met with a group of cab drivers and called ten group owners on SMS GupShup. Our outreach activities led to media coverage by both domestic and international outlets [41, 42]. The basic rules for the contest were explained in the digital promotional material and personal conversations. Internet users could also visit the contest website [43] for more detailed examples.

## 4 Analysis Methodology

To understand the results of the challenge, we employed a mix of quantitative and qualitative methods. We kept electronic logs of all calls and SMS's submitted to our server, and analyzed the approximate geographic origin of calls using the prefix of the telephone number [37]. On the qualitative side, we conducted structured phone

interviews with 84 participants, probing themes such as how they came to learn about the challenge, who they told and how they communicated about it, and what was their strategy (if any) to win. The interviews were conducted in English and Hindi by the first author (male, age 28). Each phone interview lasted around 15 min. We took detailed notes during the interview and used open coding to analyze the data. Of the 84 people we interviewed, 65 were students, 17 were employed in a private job, and 2 were homemakers. The specific participants interviewed were 31 people (of 32 participants) who submitted all five phrases; 1 person (out of 2) who submitted 4 phrases; 6 people (out of 6) who submitted 3 phrases; 38 people (out of 53) who received one of the rare phrases (where, when, or why); and 8 other participants.

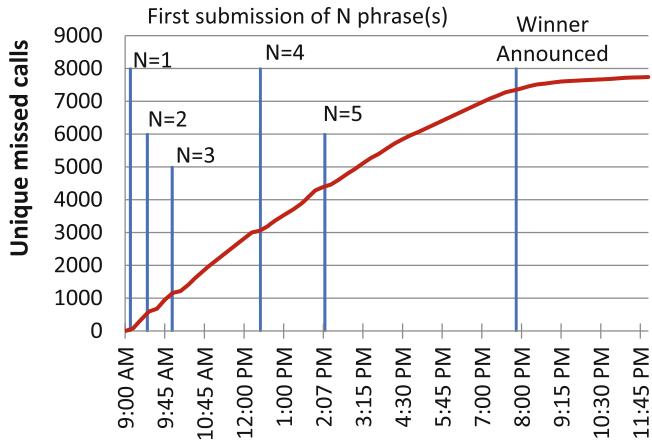
At the end of the challenge, we also invited participants to complete a brief online survey. We publicized the survey via SMS and also on the contest website, and received about 300 responses in one day. Many questions in the survey were optional and thus, different questions were answered by a different number of users. There were 167 male and 46 female respondents. The average age of the respondents was 21.4 years (s.d. = 6.28). The respondents were from 42 universities and 5 organizations. Respondents included 174 students, 14 salaried employees, 2 professors, and 1 homemaker. The majority of the users had a feature phone or basic phone. Fifty-nine respondents heard about the challenge through an email sent by a friend, college authorities or professors, 58 heard through offline conversations with friends, relatives, professors and colleagues, 47 got the information through Facebook and websites, and the remainder heard about the challenge through text messages, offline promotional events, advertisements, and tasks on Amazon Mechanical Turk. Most respondents, 192, received *Who*, 27 received *What*, 4 received *Where*, 2 received *Why* and none received *When*. Sixty-one respondents reported discovering one phrase while 65, 24, 11 and 36 participants reported discovering two, three, four and five phrases respectively. Eleven respondents could not even begin their campaign as the challenge finished much earlier than they expected. On an average, each person reported sharing their phrase with 33 people (s.d. = 120) and receiving a phrase from 30 people (s.d. = 93).

## 5 Results

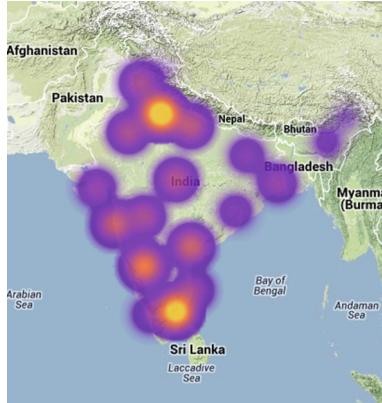
The Whodunit Challenge was launched on February 1, 2013 at 9:00 AM local time. The challenge drew 7,739 participants in less than 15 h (see Fig. 3). The first winning submission was made in just over 5 h. However, we delayed announcing that the contest was over until the evening, as we also wanted to rank and recognize the runner-up teams.

Participants sent a total of 10,577 missed calls to the system. Of the unique callers, 6,980 received the phrase for “Who”; 740 received “What”; 18 received “Where”, 17 received “When” and 17 received “Why”.

There were 185 people who submitted at least one phrase. The first person to submit two phrases did so within 26 min; 3 phrases, within 57 min; 4 phrases, within 3 h and 19 min; and five phrases (winning the contest) after 5 h and 7 min. Geographically, participation spanned across all of India, as illustrated in Fig. 4.



**Fig. 3.** Number of unique missed calls vs. time



**Fig. 4.** Heat map of received missed calls

## 5.1 Winning Strategies

The winning teams are listed in Table 1. The table lists all 20 teams who submitted 3 or more phrases and, to the best of our knowledge, discovered these phrases without help from other teams. While we are certain about the rank ordering of the first two teams, there is a complication in ranking the remaining teams: the winning team posted all of the phrases on the Facebook page of Whodunit Challenge at 4:30 pm. Thus, we rank teams by two criteria: first, by the number of phrases they submitted in advance of 4:30 pm, and second, by the total number of phrases they submitted and claimed (during our interview) to have found independently. While 13 teams claimed to have found all the phrases on their own, only 2 teams found all phrases in advance of the leak.

**Table 1.** Top 20 teams in the Whodunit Challenge

Team Number	Affiliation	Phrases submitted by 4:30pm*		Total Phrases Submitted *	Time submitted last phrase	Face-to-face	Voice Call	SMS	WhatsApp	Email	Social media	Benefactors of Prize	Notes
1	IIT Delhi (1)	5	5	2:07 PM		✓	✓	✓			✓	published incentive scheme	used SMS server; Facebook group of 474
2	IIT Delhi (1)	5	5	2:14 PM		✓	✓				✓	team leaders only	mostly used voice, SMS to reach to friends & family
3	IIT Delhi (2)	4	5	5:00 PM							✓	published incentive scheme (see text)	website with 200 registrations; FB event with 392 replies
4	Jansons Inst. of Tech.	4	5	7:00 PM		✓	✓				✓	shared with team (details unclear)	50% reached via SMS; voice; 50% via FB
5	Paavai Eng. College	4	4	3:30 PM		✓	✓	✓			✓	team leaders only	2 leaders managed 7 sub-teams of 15-20 each
6	IIT Delhi (2)	3	5	7:05 PM		✓					team leaders only	leaders focused on different geographies	one-person team; calls & WhatsApp worked best
7	IIT Delhi (3)	3	3	2:10 PM		✓	✓	✓	✓	✓		\$180-\$270 given to leaders, who distribute to sub-teams	one-person team; calls & WhatsApp worked best
8	IIM Indore	3	3	2:24 PM		✓	✓	✓				team leaders and out-of-state champions	focused on calls & SMS
9	Delhi Univ.	3	3	3:18 PM		✓	✓				✓	team leaders and out-of-state champions	focused on calls, as many do not read SMS
10	VIT Chennai	2	5	7:07 PM			✓					mostly leaders; small share (TBD) with team	used SMS exclusively
11	UPEI	2	3	PM	5:11	✓	✓	✓				team leader only	one-person team
12	LBS Institute	2	3	5:43 PM		✓						team leaders were classmates	relatives in hometown spread info to many
13	MIT Manipal	0	5	4:59 PM				✓			team leaders only	team leaders only	had classmates make two calls: local/home SIM
14	Chandigarh	0	5	5:45 PM		✓	✓	✓				team leaders only	main team leader is junior in high school
15	IIM Ahmedabad	0	5	6:01 PM		✓					✓	team leaders only	leaders asked friends to contact friends at home
16	Class 11 students	0	5	6:54 PM			✓					team leaders only	made voice calls to explain contest purpose
17	Amrita School of Engineering	0	5	7:00 PM				✓			✓	sponsor industrial visit for college	70% reached via FB; 30% via calls and SMS
18	VIT Chennai (2)	0	5	7:48 PM		✓	✓			✓		promised party for team	
19	VIT Chennai (3)	0	5	7:54 PM		✓	✓			✓		promised party for team	
20	Unknown	0	4	5:32 PM	†	†	†	†	†	†	†		†

\*We asked teams to report the total number of phrases that they submitted without help from other teams.

† Data not available

The winning team was based at the Indraprastha Institute of Information Technology Delhi (IIIT Delhi), led by 2 Ph.D. students and 6 undergraduates. In advance of the contest launch, this team set up a website<sup>3</sup> and a Facebook group<sup>4</sup> that attracted 474 members. The website publicized the following financial incentives. If the team won, they would award Rs. 10,000 (USD 167) to anyone who sent them a new phrase; Rs. 2,000 (USD 33) to anyone who directly referred someone who sent a new phrase, and a mobile top-up worth Rs. 50 (USD 0.83) to the first 200 people who sent any phrase. They set up an SMS server to which people could forward phrases. They recruited team members using a variety of methods, spanning phone calls, SMS, WhatsApp and social media platforms.

The second-place team was based at the Indian Institute of Technology Delhi (IIT Delhi), led by eight second year Computer Science undergraduates. This team finished just 7 min behind the leader. Yet they used a very different strategy: they set up a small call center, relying mostly on direct calls and SMS to reach out to family and friends who live in smaller towns and villages across the country. In turn, they asked these contacts to gather team members from the local community. One team member also set up a Facebook group and utilized Facebook group chat.

The third-place team was also based at IIT Delhi, led by six undergraduate students. This team found 4 phrases in advance of 4:30 pm, and claims to have found the fifth phrase (working independently) by 5:00 pm. Unlike other teams, this team relied solely on social media and email to recruit members. They invited over 4,000 people to a Facebook event,<sup>5</sup> of whom 329 replied with “Going” and 63 replied with “Maybe”. The group page was linked to another website where team members could register and receive a unique ID, which could be used to refer others to the team. Participation by those referred led to modest payments to the referrer (Rs. 100, or USD 1.67, for 20 referrals).

The fourth-place team, based at the Jansons Institute of Technology in Coimbatore, was led by a single undergraduate student. She estimated that she reached out to 250-300 people, half via SMS and voice calls, and half via Facebook. She submitted the fourth phrase at 3:30 pm and the fifth at 7:00 pm. While she expressed interest in sharing the prize money with team members, she did not have any incentive structure in place and the terms were not discussed with the team members; her team members helped her as a personal favor rather than for a monetary incentive.

The fifth-place team was based at Paavai Engineering College in Tamil Nadu, led by two cousins. They managed seven sub-teams with 15-20 people per team and used face-to-face interactions, phone calls, SMS, and social networks to coordinate. Interestingly, they also contacted a relative who worked at a mobile shop; the shop asked customers to give a missed call on the contest number and forward phrases to him, which he then shared with the team leaders. They did not have a formal incentive strategy, though as they got closer to winning, they offered to share a prize with those who helped them.

---

<sup>3</sup> <http://muc.iiitd.edu.in/whodunit/>

<sup>4</sup> <https://www.facebook.com/groups/528552907178873/>

<sup>5</sup> <https://www.facebook.com/events/124800261025377/>

## 5.2 Emergent Themes

Rather than describe additional teams in detail, we present three high-level themes that emerged across the remainder of our analysis. This draws from our interviews with teams, our interviews with recipients of rare phrases, and the web-based follow-up survey.

**Internet but also SMS, Voice, Face-to-Face.** All of the top five teams (and 14 of the top 19) utilized the Internet to their advantage. The most common uses were to establish a website (either independently or as a Facebook page) and to reach out to friends and contacts via Facebook (10 teams), WhatsApp (6 teams) and email (2 teams).

At the same time, teams also demonstrated a heavy reliance on non-Internet technologies: thirteen teams utilized SMS and eleven utilized voice calls to mobilize people. There were nine teams that utilized all three technologies: SMS, voice calls, and the Internet. Only three teams relied on Internet technologies alone.

The prevalence of communications outside the Internet is confirmed by our interviews with those who received rare phrases. Most often, they heard about the contest via an SMS ( $n = 9$ ) or face-to-face interaction ( $n = 8$ ) with a friend. Learning about the contest from Facebook was less common ( $n = 4$ ). Other ways of learning about the contest included email, phone calls, WhatsApp, etc. Our online survey revealed that even among participants that have access to various Internet-based services, 43 % heard about the contest through offline personal conversations with friends and colleagues.

An example of effective use of non-Internet technologies is the runner-up team (IIT Delhi), who relied mainly on a call center approach to reach family members in rural India. As a team leader explained to us, “My mom doesn’t use Internet”, and neither does the majority of India’s rural population, which constitutes 72 % of the overall population. As another example, a team of office drivers used only voice calls to manage a team and found two phrases in less than two hours.

One enterprising undergraduate (from the Amrita School of Engineering, Coimbatore) claims to have built a team of 200 peers using face-to-face contact alone. He estimates that with the help of his team’s combined efforts, he reached out to at least 1,000 people. While he reports finding three phrases, he did not submit them via SMS because he thought more phrases would be released later. (For this reason, his team does not appear in Table 1.).

These teams’ success illustrates that it is also possible to mobilize a sizable crowd without broadcast technologies such as social media or Internet websites. This finding has implications for social mobilization in areas lacking Internet connectivity, or for well-connected areas that experience Internet outages during crises.

**Reliance on Personal Networks.** In previous social mobilization contests, many teams incentivized strangers to join them. However, in the Whodunit challenge, a common thread amongst many teams’ strategies was a reliance on personal networks: team leaders reached out to their friends and family, as opposed to incentivizing lesser-known acquaintances or strangers to join their team. This is already evident in the strategies for teams 2, 4, and 5, described above, as well as for many other teams. This trend is also corroborated by the online survey where 63 respondents reported relying on friends and colleagues for discovering phrases rather than 16 respondents

who incentivized strangers. We also collected anecdotes where participants simply borrowed their friends' phones and gave the missed call on their behalf, without even explaining that there was a contest. If a new phrase was received on the friend's phone, it would be forwarded to the participant's phone. Three recipients of rare phrases reported that their phone was borrowed and used in this way. One recipient of a rare phrase was a vegetable seller who had absolutely no knowledge about the contest; we hypothesize that his phone was borrowed without offering him any explanation.

**Most Participants not Driven by Cash Rewards.** Building on the prior theme, the primary motivation for most participants was a desire to help a friend or family member, rather than any desire for (or even knowledge about) a cash award. Of the top 19 teams, less than half had any plans to distribute the cash prize beyond the inner circle of team leaders; even the runner-up team did not offer any financial incentive to its members. In teams that did plan to distribute the prize, the majority were very vague about how they might reward their full team. In contrast to the challenges conducted in developed countries, team members were motivated by non-financial factors, and any reward offered to them would be perceived more as a courtesy than as an owed compensation for their services.

We can quantify this tendency based on our interviews with those who received a rare phrase. Of the 35 respondents, only about one quarter (9) said that they were told about any financial incentive in relation to the contest. The majority (18) were not told about incentives, while the remainder (8) were team leaders or individuals who were working alone. Of the people who were not told about any incentive scheme, the majority (12/18) nonetheless shared their phrase with others. This fraction is not significantly different from those who shared their phrase with knowledge of an incentive scheme (7/9).

Some team leaders offered non-monetary incentives for their members. The leader of a team from Amrita School of Engineering, Coimbatore (#17 in Table 1) promised to sponsor a forthcoming industrial visit for his class if they won the challenge. We talked to four team leaders who proposed to throw a party for their friends if they were the winner.

Participants sometimes had intrinsic motivation to participate. For example, one student who received a rare phrase and forwarded it to benefit the winning team (IIIT Delhi) remarked, "I knew about the incentive model. Money was not important. I wanted my institute to win." An Infosys employee chose to participate and forward a rare phrase to a friend's team because he thought the contest itself was creative and worthy of participation. He thought that the purpose of the contest was to understand the role of technology in solving crime.

Some teams also experimented with other incentives. One team (not shown in Table 1) approached a charitable foundation, asking them to help publicize their team in exchange for 75 % of the prize money. While the foundation was receptive, it did not promptly post about the challenge on its Facebook page (which has over 20,000 likes) and thus offered little of the anticipated help within the timespan of the contest.

## 6 Discussion and Recommendations

While the Whodunit Challenge was quite successful in attracting enthusiastic participants from across India, the lessons learned can also serve as design recommendations to help future crowd mobilization challenges to reach out to a larger number of people, especially in low-income or offline environments.

One of the shortcomings of the Whodunit Challenge was the low level of engagement by low-income low-literate populations, primarily because we did not promote the contest widely in offline environments. The contest and the prize money appeared to be too good to be true for many low-income people that we interacted with. Many of them were uncertain about the reasons for awarding a high monetary prize just for sending missed calls. Despite our explanations, they had reservations about whether they would be charged for sending a call to our system. They were also concerned whether we would misuse their number, e.g., by sending them pesky voice calls or text messages.

To encourage more participation by non-Internet users, one approach would be to restrict promotions to offline audiences, limiting the visibility to Internet users. Another approach would be to partner with local organizations that work closely with low-income groups, distribute graphic pamphlets in local languages, and conduct outreach efforts led by people who are from the target community or have similar socio-economic status. It could also help to make the contest harder, for example, by decreasing the frequency of certain phrases or enforcing geographical diversity of team members (in India, coarse-grained geographic information can be determined from the caller ID. [44]). As teams are forced to reach out to broader populations, they may derive greater benefit from reaching out to the masses of rural and lower-connectivity residents. Disseminating phrases in audio format rather than text would also enable inclusion of lesser-educated participants, though the cost of phone calls could be a significant deterrent (either for participants or for the challenge organizers, depending on who pays for the calls.)

One of our interesting findings is that participants were often motivated by non-monetary incentives, including social support for friends and recognition for their institution. Future challenges might employ non-monetary incentives to increase participation, for example, by offering recognition, goods or services that cater to groups (such as a party or travel vacation).

Our usage of mobile phone numbers as a unique personal identifier was largely successful in prompting the formation of large teams. However, it also led to some subtle implications, such as the practice of borrowing others' phones to leverage their participation without their full knowledge or consent. While we did not observe any serious abuses of this situation, e.g., by stealing phones or feeding misinformation to potential participants, these possibilities are nonetheless important to consider and guard against in future challenges.

One limitation in the design of the Whodunit Challenge is that it is not possible to know the exact sizes of teams. Addressing this limitation would have required a fundamental change in the contest dynamics, for example, to require each participant to identify themselves with one or more teams. This would likely require an interaction

richer than a missed call, which would have added cost and complexity for participants. Though sending SMS may seem easy, only 185 of 7,739 participants submitted a phrase to our server. Some participants may have been motivated only to share their phrase with their team leader, while other participants may have had limited familiarity with SMS and how to forward them. In any case, finding creative techniques to more accurately track the growth and composition of teams, without adding complexity for participants, could yield large benefits in the analysis of future challenges. One potential approach could be to host a large number of contest phone numbers, each advertised to a small number of people. If two participants place calls on different numbers, it would be unlikely that they are on the same team.

Our final recommendation is to take extra care in designing simple rules and communicating them to participants. Though we distilled the challenge rules to very simple language, including several illustrative examples, many teams misunderstood aspects that prevented them from competing well. We found three teams who thought that some phrases would be released at a later date, preventing them from being aggressive in the initial stages. We talked to five teams who assumed that phrases would be distributed across different geographical regions, causing them to seek out more geographies rather than seeking out more people. We also spoke with five teams who assumed that all phrases needed to be submitted together, preventing them from gaining feedback and recognition for intermediate progress. It is important to anticipate any possible misconceptions and proactively convey the requisite clarifications. Several individuals misunderstood each phrase to be a puzzle instead of a part of the secret sentence; for example, in response to “Who: Rajnikanth”, they would respond with “actor”. While these details are somewhat specific to the Whodunit Challenge, the broader implication is that though it is difficult, it is necessary to design simple rules that are easily understood and easily communicated from one person to another. This is especially important for lesser-educated participants and those who may lack the devices, connectivity or bandwidth to view large explanatory materials (such as websites, promotional videos, etc.). We also recommend setting up more accessible information portals, such as an Interactive Voice Response system, to make the rules more accessible for people with low literacy and limited access to the Internet.

## 7 Conclusion

This paper presents the first crowd mobilization challenge conducted in India, a developing-country context where effective social mobilization is broader and more inclusive than the rich-country settings studied previously. We customized the design of the challenge to incorporate local languages and to enable participation at very low cost by anyone with access to a basic mobile phone. The challenge was successful in attracting broad participation, spanning 7,700 participants from all across India in less than a day. While many participants utilized Internet technologies, we also found interesting usage of SMS, voice, and face-to-face communications that offered benefits in the Indian context. Unlike previous social mobilization contests, participants relied primarily on their personal networks, and often recruited team members without offering any financial incentives. We synthesize our lessons learned as a set of

recommendations to help future crowd mobilization challenges extend their reach into low-income, offline environments.

## References

1. Von Ahn, L.: Duolingo: Learn a language for free while helping to translate the web. In: Proceedings of the 2013 International Conference on Intelligent User Interfaces, pp. 1–2. ACM, New York, NY, USA (2013)
2. Bigham, J.P., Jayant, C., Ji, H., Little, G., Miller, A., Miller, R.C., Miller, R., Tatarowicz, A., White, B., White, S., Yeh, T.: VizWiz: Nearly real-time answers to visual questions. In: Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology, pp. 333–342. ACM, New York, NY, USA (2010)
3. Chen, K., Kannan, A., Yano, Y., Hellerstein, J.M., Parikh, T.S.: Shreddr: Pipelined paper digitization for low-resource organizations. In: Proceedings of the 2nd ACM Symposium on Computing for Development, pp. 3:1–3:10. ACM, New York, NY, USA (2012)
4. Hara, K., Le, V., Froehlich, J.: Combining crowdsourcing and google street view to identify street-level accessibility problems. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 631–640. ACM, New York, NY, USA (2013)
5. DARPA Network Challenge. <http://archive.darpa.mil/networkchallenge>
6. Pickard, G., Pan, W., Rahwan, I., Cebrian, M., Crane, R., Madan, A., Pentland, A.: Time-Critical social mobilization. *Science* **334**, 509–512 (2011)
7. Rahwan, I., Dsouza, S., Rutherford, A., Naroditskiy, V., McInerney, J., Venanzi, M., Jennings, N.R., Cebrian, M.: Global manhunt pushes the limits of social mobilization. *Computer* **46**(4), 68–75 (2013)
8. MyHeartMap Challenge. <http://www.med.upenn.edu/myheartmap/>
9. Wilcock, D.: Police Facewatch app targets London riot suspects. <http://www.independent.co.uk/news/uk/crime/police-facewatch-app-targets-london-riot-suspects-7887778.html>
10. Crowdsourced videos, photos could aid Boston blast investigations. <http://www.cnet.com/news/crowdsourced-videos-photos-could-aid-boston-blast-investigations/>
11. Kaufman, L.: Bombings Trip Up Reddit in Its Turn in Spotlight, (2013). <http://www.nytimes.com/2013/04/29/business/media/bombings-trip-up-reddit-in-its-turn-in-spotlight.html>
12. Wemple, E.: Young men, please sue the New York Post, (2013). <http://www.washingtonpost.com/blogs/erik-wemple/wp/2013/04/22/young-men-please-sue-the-new-york-post/>
13. Telecommunications in India, (2015). [http://en.wikipedia.org/w/index.php?title=Telecommunications\\_in\\_India&oldid=656139185](http://en.wikipedia.org/w/index.php?title=Telecommunications_in_India&oldid=656139185)
14. Census of India - Mode of Communication: 2001–2011. [http://www.censusindia.gov.in/2011census/hlo/Data\\_sheet/India/Communication.pdf](http://www.censusindia.gov.in/2011census/hlo/Data_sheet/India/Communication.pdf)
15. GupShup Re-launches GupShup Messenger, Innovates for India. <http://www.siliconindia.com/news/technology/GupShup-Reaunches-GupShup-Messenger-Innovates-for-India-nid-140878-cid-2.html/1>
16. Indian Railways, (2015). [http://en.wikipedia.org/w/index.php?title=Indian\\_Railways&oldid=657313685](http://en.wikipedia.org/w/index.php?title=Indian_Railways&oldid=657313685)
17. Kumaraguru, P., Cranor, L.F.: Privacy in india: attitudes and awareness. In: Danezis, G., Martin, D. (eds.) PET 2005. LNCS, vol. 3856, pp. 243–258. Springer, Heidelberg (2006)
18. Kittur, A., Nickerson, J.V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., Lease, M., Horton, J.: The future of crowd work. In: Proceedings of the 2013 Conference on Computer Supported Cooperative Work, pp. 1301–1318. ACM, New York, NY, USA (2013)

19. Tang, J.C., Cebrian, M., Giacobe, N.A., Kim, H.-W., Kim, T., Wickert, D.: Beaker: reflecting on the DARPA red balloon challenge. *Commun. ACM* **54**, 78–85 (2011)
20. Dodds, P.S., Muhamad, R., Watts, D.J.: An experimental study of search in global social networks. *Science* **301**, 827–829 (2003)
21. Mason, W., Watts, D.J.: Financial Incentives and the Performance of Crowds. In: *Proceedings of the ACM SIGKDD Workshop on Human Computation*, pp. 77–85. ACM, New York, NY, USA (2009)
22. Cebrian, M., Coviello, L., Vattani, A., Voulgaris, P.: Finding red balloons with split contracts: robustness to individuals selfishness. In: *Proceedings of the Forty-fourth Annual ACM Symposium on Theory of Computing*, pp. 775–788. ACM, New York, NY, USA (2012)
23. Naroditskiy, V., Rahwan, I., Cebrian, M., Jennings, N.R.: Verification in referral-based crowdsourcing. *PLoS ONE* **7**, e45924 (2012)
24. Ford, C.M.: Twitter, Facebook, and Ten Red Balloons Social Network Problem Solving and Homeland Security. *Homeland Security affairs*, USA (2011)
25. Rutherford, A., Cebrian, M., Dsouza, S., Moro, E., Pentland, A., Rahwan, I.: Limits of social mobilization. *Proc. Natl. Acad. Sci.* **110**, 6281–6286 (2013)
26. Wired Magazine's Vanish Challenge. [http://www.wired.com/vanish/2009/11/ff\\_vanish2](http://www.wired.com/vanish/2009/11/ff_vanish2)
27. Eagle, N.: txteagle: mobile crowdsourcing. In: Aykin, N. (ed.) *IDGD 2009. LNCS*, vol. 5623, pp. 447–456. Springer, Heidelberg (2009)
28. Gupta, A., Thies, W.: mClerk: enabling mobile crowdsourcing in developing regions. In: *CHI* (2012)
29. Ipeirotis, P.: Demographics of Mechanical Turk. Presented at the NYU Center for Digital Economy Research Working Paper CeDER-10–01 March (2010)
30. Ross, J., Irani, L., Silberman, M.S., Zaldivar, A., Tomlinson, B.: Who are the crowdworkers?: shifting demographics in mechanical turk. In: *CHI 2010 Extended Abstracts on Human Factors in Computing Systems*, pp. 2863–2872. ACM, New York, NY, USA (2010)
31. Khanna, S., Ratan, A., Davis, J., Thies, W.: Evaluating and Improving the Usability of Mechanical Turk for Low-income Workers in India. In: *Proceedings of the First ACM Symposium on Computing for Development*, pp. 12:1–12:10. ACM, New York, NY, USA (2010)
32. Gao, H., Barbier, G., Goolsby, R.: Harnessing the crowdsourcing power of social media for disaster relief. *IEEE Intell. Syst.* **26**, 10–14 (2011)
33. Hester, V., Shaw, A., Biewald, L.: Scalable crisis relief: crowdsourced sms translation and categorization with mission 4636. In: *Proceedings of the First ACM Symposium on Computing for Development*, pp. 15:1–15:7. ACM, New York, NY, USA (2010)
34. Wyche, S.P., Forte, A., Yardi Schoenebeck, S.: Hustling online: understanding consolidated facebook use in an informal settlement in nairobi. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2823–2832. ACM, New York, NY, USA (2013)
35. Rangaswamy, N., Cutrell, E.: Re-Sourceful networks: notes from a mobile social networking platform in india. *Pac. Aff.* **85**(3), 587–606 (2012)
36. Desai, S.B., Dubey, A., Joshi, B.L., Sen, M., Shariff, A., Vanneman, R.: *Human Development in India: Challenges for a Society in Transition*. Oxford University Press, New Delhi (2010)
37. Meeker, M., Wu, L.: Internet Trends, KPCB (2014)
38. Internet Live Stats. <http://www.internetlivestats.com/internet-users/>
39. Poverty headcount ratio at \$2 a day (PPP) (% of population). <http://data.worldbank.org/indicator/SI.POV.2DAY>

40. Lee, P.: From Computing Research to Surprising Inventions
41. Microsoft's social Whodunit competition to begin in India. <https://in.news.yahoo.com/microsofts-social-whodunit-competition-begin-india-045205972.html>
42. Social whodunnit competition launches in India - tech - 31 January 2013 - New Scientist. <http://www.newscientist.com/article/mg21729025.500-social-whodunnit-competition-launches-in-india.html>
43. The Whodunit? Challenge by Microsoft Research India. <http://www.whodunitchallenge.com/>
44. Mobile telephone numbering in India, (2015). [http://en.wikipedia.org/w/index.php?title=Mobile\\_telephone\\_numbering\\_in\\_India&oldid=657910549](http://en.wikipedia.org/w/index.php?title=Mobile_telephone_numbering_in_India&oldid=657910549)

# Wayfinding Behavior in India

Naveed Ahmed<sup>(✉)</sup>

Industrial Design Centre, Indian Institute of Technology Bombay, Mumbai, India  
ahmed.naveed@yahoo.co.in

**Abstract.** India is highly heterogeneous in the way cities are laid out; the way people use landmarks and how certain cultural aspects affect wayfinding. These factors influence the design of maps and navigation systems.

Users from Mumbai and Rajasthan were interviewed to explore such implications and find design opportunities. These users had low education levels and needed to find addresses on a regular basis. The study was centered on the Indian context.

People in India rarely use maps for navigation. They rely primarily on asking around and navigate using landmarks. The landmarks people use during this are prominent but sometimes highly volatile and ad hoc like cows and people sitting on street corners. Some of these landmarks may not necessarily always be popular. While inquiring about the route, people repeatedly seek reliable sources en route, to validate the information they have. Other findings during the study include people's preferences in using maps and concerns while seeking directions. Mental models of people also affect the way people navigate and exchange the wayfinding information. Some of these are very specific to the Indian context.

In the end, we also discuss how these findings will affect the design of navigation and (culture-centric) wayfinding systems.

**Keywords:** Navigation · Wayfinding · India · Culture · Behavior

## 1 Introduction

Navigation and wayfinding are an integral part of modern lives. Wayfinding is the process of gathering required information and making decisions about the direction of travel. This includes determining one's location in comparison to the destination or a nearby location.

According to Golledge, wayfinding is concerned principally with how the route is structured rather than the environment through which the path passes [1], but in a modern city where life is affected by factors like traffic congestion, multitasking, etc., the path and the environment equally impact wayfinding and navigation.

The process of wayfinding is affected by a person's physical, intellectual abilities, socio-cultural background, gender, etc. [2, 3] and in turn adversely affects the design of wayfinding systems. In India, where people rely primarily on asking around, these factors even determine if the information conveyed is right or wrong. People have apprehensions and preferences in asking around. People take time to decide whom to ask, when to ask and how often to ask.

Along with these, mental models and perceptions of a particular place determine how people store information about routes and how they are conveyed to other people. As navigation is a spatiotemporal process, these perceptions change based on the stimuli the traveler receives at the moment [4]. This results in creation of newer wayfinding strategies and routes, than that already exist [5].

India as a country is highly heterogeneous in the way cities are laid out and the way people use landmarks in the absence of proper road signage. Cultural aspects specific to this country determine how wayfinding instructions are exchanged. These affect the way maps and navigation systems are designed. The studies reported in this paper are an attempt to explore such implications and find design opportunities.

In a map, landmarks like buildings, trees, statues, etc. and topographical features like hills and elevations act as referential points to know one's position or to orient oneself. These are marked on a physical map to aid navigation. But in India, where very few people use printed maps, landmarks play a pivotal role and wayfinding is done by asking around for these landmarks en route. Added to these are many biases of how people provide these instructions that affects the process of wayfinding.

The findings of user studies have been presented further and possible implications that affect the design of wayfinding systems in a country like India are discussed.

## 2 Methodology

As part of a Master's project, we tried to understand people's wayfinding habits, how people use existing solutions and conditions under which these have failed. The findings in this paper are the result of user studies conducted during the project.

To understand the process of wayfinding, users who did this on a regular basis were interviewed. This included delivery personnel, intra- & inter-city drivers and repair technicians. Most of the users had low education levels and needed to find addresses on a frequent basis. People who used existing solutions regularly were also interviewed to understand their usage behavior.

Users were chosen based on their vocation and availability. In all, 15 people were interviewed and most of them were in the age group of 25–35 and have lived in city suburbs most of their lives. Of these, 10 were men and 5 women. Finding unique users was time-consuming as these cohorts operate within fixed areas in the city. The users had low familiarity with using a computer but used smartphones regularly.

The project also took into account and extended the outcome of a study on wayfinding by design students at IIT Bombay, which involved interviews with 47 people in towns near Mumbai.

Indian cities are structured differently and hence users were chosen from geographically different regions to obtain diversity in inputs—two of which were inter-city drivers from Jaipur (Rajasthan) and the rest from Mumbai and its suburbs.

Contextual inquiries were conducted with the users to know their working processes, how they find addresses, their recent journeys and the hurdles they face. As part of unstructured studies, something similar to the master-apprentice model was followed. We chose a few destinations and asked random people (en route) for ways to reach the

place. In both the cases, the responses were recorded and later analyzed through affinity mapping.

Towards the end of the primary studies, a separate set of 5 people were interviewed to understand mental models of sketching and representation of familiar routes.

The project resulted in the conceptualization of a navigation mechanism using schematic maps and photographs of landmarks. This has not been discussed in this paper.

### 3 Findings

The findings of the user studies, discussed below, affect the way maps are designed, particularly in the Indian context.

#### 3.1 Landmark Information

As India is not a very map-savvy culture, landmarks have always played an important role in wayfinding. In our studies, even users of online maps mentioned that they checked major landmarks on their route before starting the journey.

Prominent physical structures like buildings and trees are commonly used as landmarks. Sometimes the color of a building is more important than the building itself. But this is mainly in case of larger roads.

In smaller streets, the choice of landmarks changes completely. People use garbage bins, broken walls, *nullahs* and even volatile entities like road-side vendors, people sitting on street corners, potholes and cows as reference points. Here, religious structures like minarets and *nishan sahib* which are specific to an area are also used.

**Some landmarks from user studies (specific to India):** Mother dairy outlets (milk booths), trees, color of buildings, slums, electric poles, banyan trees, fish markets, post-boxes, tiny temples and *dargahs* under trees, rickshaw stands, road-side shops (like ironing store, cobblers, cycle shop, betel nut & cigarette shops), street vendors, drainages, broken walls, garbage bins, *chabutras*, political party offices, police outposts, policemen, gated communities (like *chawls*), minarets, electric transformers, religious flagpoles (e.g. *nishan sahib*), prominent cultural residences (like *pols* in Ahmedabad), public toilets, grocery stores, flyover pier numbers are amongst other prominent locations.

This list is not exhaustive and is in no specific order.

#### 3.2 Whom to Ask?

Though people rely heavily on asking around to find the address right, they are often not sure about whom to ask for directions. People solve this problem by relying on ‘seemingly-trusted’ persons and ‘possible’ providers of information.

Well-dressed and older people are presumed to be more trustworthy than others. *Autowallahs* and *panwallahs*<sup>1</sup> are considered more reliable as they are usually local to an area and might know the place better. But some users said that *autowallahs* often refused or provided wrong information.

A user also mentioned a case where school students misguided him and he had to find his route all over again.

Women prefer asking other women. This striking feature was noticed more than once in our studies. According to Lawton and Kallai, women are more likely to experience wayfinding anxiety than men. This could be due to the cultural upbringing resulting in a feeling of insecurity and vulnerability, especially in unfamiliar areas. Studies have also shown that women use different wayfinding strategies than men [3].

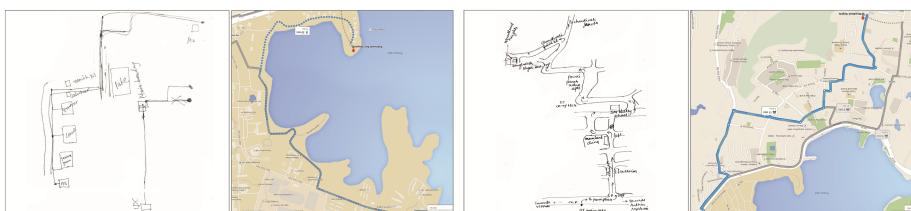
People also seek multiple sources (en route) to verify and validate the information they have been provided with. But this may sometimes cause more harm than good, as people have their own perceptions and understanding of a route and excessive verification by the information seekers (info-seekers) might lead them astray.

### 3.3 Mental Models

People create mental models and have perceptions in navigation and wayfinding too. Users construe a route without any turns as a straight road. They do not perceive a curve unless it is at a sharp angle. Users subconsciously perform rectilinear correction of the path by ignoring the minor bends and curves.

When asked to draw the route between two points they travel regularly, users usually did a correction in the path. If the route had no major turns, it was drawn as a straight road. Bends were ignored. People mentioned details only when the path involved too many turns. They sketched finer details only when they got closer to the destination.

High-speed parts of the road were drawn straighter than the low-speed ones (Fig. 1).



**Fig. 1.** On the left are sketches by people and on the right are screenshots from Google Maps.

However, the nature of sketch might change depending upon how familiar the route is to the person. Regular commuters, infrequent travelers and first-time travelers will have a different perception of the same route. Understanding the difference in their mental models will require further study and inquiry into this subject.

<sup>1</sup> An autorickshaw driver is a *autowallah* and a *panwallah* is a betel nut seller. Both of these can be found in most of the street corners in India.

Kevin Lynch in his book, *The Image of the City*, talks about how landmarks should be created in a city so that people can create easy mental maps of the city they live in and of the routes they traverse (called imageability and visibility) [6]. So, these mental models are also affected by the layout and structure of the city.

For example, cities like Chandigarh in India follow a grid (also called gridiron) structure, while New Delhi has a radial structure and places like Hyderabad which have evolved organically, do not follow any fixed structure. As grid cities are organized in rectilinear roads, referring to and recalling routes is easier, than in case of an organic or non-grid layouts.

Roads look different at different times of the day. Landmarks also appear different because of traffic and light conditions which vary greatly depending on the time of the day. The perception of a route, hence, also depends on several natural and man-made variables.

### 3.4 Progressive Discovery

People find their way through progressive discovery. Progressive discovery is like progressive disclosure which is an interaction design technique where only the important and primary information is disclosed initially and the rest of it is provided on demand [7]. Similarly, users divide their journey into shorter milestones and once they reach each of these, they inquire for further directions.

Depending upon proximity to the destination, information providers (info-providers) also provide directions progressively. They advise the info-seekers to reach a certain point ahead and then ask further. Certain info-seekers and even info-providers felt that verbal directions are complex and confusing. Hence, they followed progressive disclosure of information intuitively. In short, both info-seekers and info-providers first search for larger information and get into finer details only when they are closer to the destination.

This is true even in case of some Google Maps users who obtained the approximate route to their destination online and got into micro details of reaching their destination by asking around.

### 3.5 Size of the Road

People feel it is easier to ask for directions on larger roads than in the inner streets. This may be because of progressive discovery. People seek major landmarks while on the larger roads and get into intricate details only when they get closer to their destination. Intricate details mean specificity and finding them is more difficult.

Info-providers would generally know the bigger picture better than the finer details, compared to inner streets where the volatile landmarks are known only to those who know the area well.

### 3.6 Measuring Distance

People use several indigenous and ingenious methods to measure distance. Locally available data such as the number of buildings, telephone poles, traffic signals and trees

are often used. In some cases, the distance is measured by the amount of fare charged by the bus or any available mode of public transport. In rural areas, people also use the number of villages between the starting point and the destination as a measure of distance.

The number of route options to reach the destination depends directly on the distance between the source and the destination. The layout of the city also affects the number of routes available. A grid structure has lesser route options to complete a journey compared to a radial or organic layout [8]. One user mentioned that grid meant more travel whereas it was easier for him to find shortcuts in a city with a non-grid layout. As a result, he perceived travel times to be lower in such (non-grid) cities.

### **3.7 People Do not Prefer Online Maps**

Being a society where maps are sparsely used for wayfinding, online aerial maps are something that users in India find difficult to adapt to. Traditional GPS-based solutions evolved from aerial perspective of locations. But in India where use of traditional maps is low, digital solutions based on aerial perspectives may not always work.

Users mentioned that online maps do not contain adequate information for wayfinding. People in India usually find their way using landmarks and by asking around. This information is not readily available to them on the maps or easily accessible. Even if they find the right route, they could not backtrack or travel again on the same route without assistance. This created a negative bias towards online maps.

Landmark information was something they wanted maps to provide, so that they can remember the route even if they are not using maps during subsequent travel.

Some users were not comfortable using software for navigation, as they could not operate them easily. Anxiety during travel and unfamiliarity with the place and the use of software may contribute to this behavior.

Users were not very keen on continuously using mobile phones for maps as it was perceived as ‘talking on the phone’ by others. A user was once booked by traffic police for using mobile phone while the vehicle was waiting at a traffic signal. Also, solutions like Google Maps did provide audio feedback but it was difficult to listen to instructions in noisy traffic conditions.

### **3.8 Usage of Existing Solutions**

Of all the users interviewed, five of them had used print maps and online solutions like Google Maps and Nokia HERE earlier.

One user (a European citizen) who used print maps extensively found it difficult to use them in India as they were not detailed and finding street names was not easy. This is due to improper and insufficient signage. He later resorted to asking around.

A user felt that Google Maps provided information that was not completely reliable. He used the application primarily to know his current position and if the taxi was taking him on the correct route. But the app always re-routed based on the current position. There was no direct way to confirm the authenticity of his route.

Due to the shortest path strategy, the map applications sometime route user via smaller streets and villages, instead of the main roads or the highways. This confused users about the authenticity of the route provided. Users in such cases had to go back to asking people around to make sure they are on the right path. It leads to further anxiety when the app re-routes to a new route in an already unfamiliar path.

One user was gifted a dashboard console for his car a year back, but had stopped using it within a month. He did not find it useful as most of the travel was intra-city and it was too cumbersome to input data to find a route.

## 4 Design Implications

All the entities that form part of the wayfinding process seem to have some sort of uncertainty associated with them—may it be the mental models, the way people exchange information or the kind of landmarks that are used. In spite of all this, the seemingly-chaotic process works successfully, and for many users most of the navigation information becomes expendable after they cross a milestone during the journey.

In the findings so far, we also see that existing solutions failed due to various internal or external reasons like interface issues, poor infrastructure or specific requirements by users. Sometimes, as the profile of the user changed, so did their requirements.

Landmarks, big or small, are the key elements in the wayfinding process especially in absence of proper signage. And for a culture that navigates primarily on verbal on-the-go basis, a solution must seamlessly integrate landmark data and provide it along the route being traversed. Instead of an aerial view, the solution must be completely landmark-based that shows a ground-level view to the user.

Also in a country like India, where topographical changes are very frequent, care must be taken that landmark data is valid and correct at any given point of time. The solution must be flexible enough to accommodate the volatility of the landmarks. This will help in making the system more reliable.

People seek trustable sources and they will look for ‘trust’ in any solution. This has been clearly noticed in the user studies where people tried to seek reliable sources repeatedly. Creating trust into a non-human entity is something that would make or break the solution. As Progressive Discovery is an almost accepted behavior, it is not required that complete information is provided to the user in one go. Necessary information can be provided to the user to take him to the next intermediate point in his journey and he can take further decisions thereafter. The solution can have smaller milestones for people to achieve, like a game, leading to the final destination. This will help the user be assured that he is on the right path and also get a feeling of accomplishment.

The kind of landmark information people use also has a major impact on the way data is collected for the solution. It would be easier to find data in the bigger context than the intricate street-level details. The details from the last mile of journey are the most important and if the solution fails at this stage it would render the whole exercise of wayfinding useless. The main aim is to create a solution that works for first-time users, as people would remember the route (even if vaguely) on subsequent travel.

There is also an opportunity to create solutions, which cater to different times of the day—mainly lighting conditions—to provide a seamless experience.

Human mind ignores minor bends and curves to create unique mental models. This feature can be incorporated in a map to provide the simplest of information, similar to a schematic map. Schematic maps are closest to mental models in terms of simplicity and have been used over the years for representing transit routes. E.g. the London Tube. Major turns and prominent topological features can be reproduced for guidance and to prevent errors. Users can concentrate on their route instead of the other details, if they are provided with simplest and only the necessary information.

It may not always be possible to create an ultimate solution. But these findings might lead to a solution for certain target groups like people with low-education levels or those unfamiliar with using maps, who amount to a fairly large and untapped part of the population in India.

## 5 Future Work

In the past few years, India has seen the emergence of professions like delivery, courier, cab and related door-to-door services. These require searching and finding addresses on a daily basis. Most of these professions do not require very high educational levels and hence people employed in these are not highly educated. (Low education levels here does not imply low literacy.)

People in these professions often possess smartphones but never use them beyond the employer-specified application. Cohorts like these, who underuse accessible technology, can be the primary focus to create a solution. This must also be done by keeping in mind the bottom of the pyramid to facilitate easy adoption amongst users with very less or no exposure to technology.

Further research would include making such applications more usable through newer technologies like text-less and aural interfaces.

**Acknowledgements.** I would like to thank my project guide Prof. Anirudha Joshi of Industrial Design Centre (IDC), IIT Bombay, for his critical feedback throughout. Faculty at IDC for their inputs during the project. Participants and friends who assisted me during the user studies. Anshumali Baruah and Ishneet Grover for the initial reviews. Umme Hani for the thorough proofreading and suggestions during the making of this document.

## References

1. Golledge, R.G.: Wayfinding Behavior: Cognitive Mapping and Other Spatial Processes. Johns Hopkins University Press, Baltimore (1999)
2. Arthur, P., Passini, R.: Wayfinding: People, Signs, and Architecture, Reprint edn. Focus Strategic Communications Inc., Canada (2002)
3. Lawton, C.A., Kallai, J.: Gender differences in wayfinding strategies and anxiety about wayfinding: a cross-cultural comparison. *Sex Roles* **47**, 389–401 (2002)
4. Montello, D.R., Sas, C.: Human Factors of Wayfinding in Navigation. CRC Press/Taylor & Francis Ltd., London (2003)

5. Hunter, S.: Spatial Orientation, Environmental Perception and Wayfinding. IDeA Center, University at Buffalo (2010)
6. Lynch, K.: The Image of the City. The MIT Press, Cambridge (1960)
7. Nielsen, J.: Progressive Disclosure, 4 December 2006. <http://www.nngroup.com/articles/progressive-disclosure/>
8. Walker, J.: The power and pleasure of grids (2010). <http://www.humantransit.org/2010/02/the-power-and-pleasure-of-grids.html>. Accessed 26 May 2014

# Evaluating Digital Tabletop Collaborative Writing in the Classroom

Philip Heslop<sup>(✉)</sup>, Anne Preston, Ahmed Kharrufa, Madeline Balaam,  
David Leat, and Patrick Olivier

Newcastle University, Newcastle upon Tyne, UK  
{philip.heslop, anne.preston, ahmed.kharrufa,  
madeline.ballam, david.leat, patrick.olivier}@ncl.ac.uk

**Abstract.** We present an evaluation of an “in the wild” classroom deployment of Co-located Collaborative Writing (CCW), an application for digital tabletops. CCW was adapted to the classroom setting across 8 SMART tables. Here, we describe the outcomes of the 6 week deployment with students aged 13–14, focussing on how CCW operated as a tool for learning within a classroom environment. We analyse video data and interaction logs to provide a group specific analysis in the classroom context. Using the group as the unit of analysis allows detailed tracking of the group’s development over time as part of scheme of work planned by a teacher for the classroom. Through successful integration of multiple tabletops into the classroom, we show how the design of CCW supports students in learning how to collaboratively plan a piece of persuasive writing, and allows teachers to monitor progress and process of students. The study shows how the nature and quality of collaborative interactions changed over time, with *decision points* bringing students together to collaborate, and how the role of CCW matured from a *scaffolding mechanism* for planning, to a tool for *implementing* planning. The study also showed how the teacher’s relationship with CCW changed, due to the designed visibility of groups’ activities, and how lesson plans became more integrated utilizing the flexibility of the technology. These are key aspects that can enhance the adoption of such technologies by both students and teachers in the classroom.

**Keywords:** Digital tabletops · Collaborative learning · Multi-touch

## 1 Introduction

Digital tabletops have been described as a collaborative tool that can impact educational processes in the classroom [7, 14]. They are considered a medium for social learning [7], and have potential to foster a more collaborative, group-based approach to learning, such as students being exposed to different viewpoints (and possible solutions), developing critical thinking skills and a more nuanced understanding [23].

Recent learning applications have been designed to take advantage of the digital tabletop medium [10, 25, 29, 36]. Much of this research can be characterised as designing applications to exploit the affordances of the digital tabletop to effectively support collaborative learning [7, 21]. For example, DigiTile [24, 25] takes advantage

of the visuospatial qualities of the tabletop to support learning about fractions. Students spatially manipulate tiles to fill a canvas to visually represent fractions. Digital Mysteries [10, 12–14] also takes advantage of the visuospatial qualities of the digital tabletop, allowing students to resize, group and connect information to help the groups collaboratively formulate an answer to a question without a distinct “correct” answer. The Collocated Collaborative Writing application (CCW) [8] builds on these design principles, producing a tool for learning extended writing collaboratively. In general, the designs have been found to successfully support collaborative learning interactions, with evaluations showing:

1. Designing for the digital tabletop is more than a remediation of content; it requires specific design in order to fully utilise their affordances [7, 10, 25].
2. Externalisation of thinking can be facilitated by the digital tabletop through visuospatial representations [8, 10].
3. Applications can regulate learning by splitting tasks into stages and through scaffolding [8, 10].

These applications have shown learning improvements for single groups but have not been evaluated in a whole-class setting.

Increasingly, work has been published describing the deployment of multiple tabletops into learning contexts in the wild. Descriptions of classroom deployments of multiple tabletops have shown that there are significant differences and difficulties not found in single tabletop deployments [13] and even in multi-tabletop scenarios that are not integrated into an authentic curriculum setting [11]. This requires:

- Sessions take place in an ordinary classroom based in a school.
- Multiple simultaneous groups working on multiple digital tabletops.
- Supervision by the teacher who usually teaches the lesson to the students.
- Teacher created content for the sessions based on their teaching goals.
- Integration of the technology into teachers’ lesson plans and tying the activity to specific learning goals.

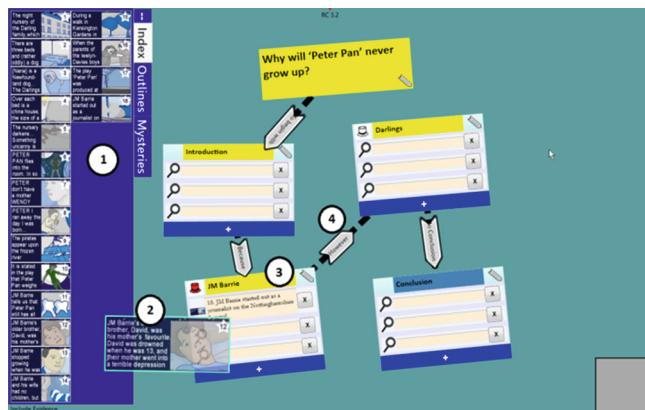
The findings have commonly focussed on the design of technological strategies that can support the teacher’s management and orchestration of the class [20]. It is clear from these earlier “in the wild” multi-tabletop deployments [11, 13] that the precursor to an adequate evaluation of the impact of the tabletops, is for teachers to embrace the technology as a tool they can use effectively to support learning by integrating it into lesson plans.

In this paper, we contribute a learning focussed evaluation of a multi-tabletop deployment in a classroom supervised by a teacher during normal lesson time. In so doing, we (1) focus primarily on the nature and quality of one group’s collaboration through the students’ communicative interactions with the technology and each other, (2) track the group’s development over time as part of scheme of work planned by the teacher, and (3) use classroom level data streams to show how this is indicative of the class as a whole. I.e. using the group as the unit of analysis to generalise the collaborative behaviours occurring in the classroom [28].

## 2 Collocated Collaborative Writing

Collocated Collaborative Writing (CCW) [8] is an application designed to facilitate the learning of extended writing in small groups by exploiting the affordances of digital tabletops. Extended Writing can be characterised as any writing using specialised vocabulary and a formal structure. Planning is an essential skill in producing high level structured writing [15, 17] and good quality plans tend to be well structured (i.e. include well-connected paragraphs), on topic (i.e. not generic or abstract) and lead to higher quality final documents [3]. Writing Frames [18] are a paper-based scaffold to support extended writing. Using specific genres, the method provides partial plan-like structures to be completed by students.

CCW builds on Writing Frames and focuses on persuasive writing (one of the more difficult genres). The genre requires the creation of a persuasive argument across several paragraphs, including supporting evidence and consideration of alternative interpretations. Unlike the paper based Writing Frames, CCW allows the document structure to be changed dynamically. Figure 1 shows the current CCW interface.



**Fig. 1.** CCW Interface: 1. Evidence Palette 2. Evidence Slips 3. Paragraphs 4. Connection

The task is split into stages, and students themselves decide (with a decision point) when to attempt to progress. If certain criteria are not met, then the students are given scaffolded instructions to help them towards completing the current stage.

The four stages are: (1) **Examine Evidence** – students read through the evidence slips. (2) **Create (Named) Paragraphs** – students create new paragraphs and give appropriate names – a minimum of four paragraphs must be made to progress. (3) **Connect Paragraphs** – to progress *all* paragraphs must be connected and (4) **Use Evidence:** where evidence slips are inserted into paragraphs – to progress, each paragraph must have a minimum of three evidence points, either from slips or created by students.

CCW is designed on principles of collaborative learning [33], incorporating ideas from distributed cognition with focus on the use of space and the manipulation of

representations [6, 16, 22, 27, 37]. It also includes teaching methods such as scaffolding [34, 35] to complement and support the scaffolding supplied by teachers. The two main interaction design concepts leveraged to turn writing into a collaborative process are: (1) the use of a visuospatial design allows for representation and communication of ideas, i.e. paragraphs, evidence, and connectors are created as visual representations that can be manipulated by multiple users allowing visual externalisation and communication of ideas. (2) The introduction of decision points throughout the process: between stages, creation of a paragraph, and adding connections. Such decision points help regulate the progress and prompt collaborative discourse, i.e. Proposals [1]. The interface is *cumulative*; that is, no functionality is lost between stages and additional functionality works on the existing state. Accordingly, the decisions made to reach the current representation can be determined by observing the current state.

The iterative design process of CCW [8] was based on single group studies. Previous integration into the classroom [13] highlighted significant differences between the requirements for a single group and the requirements of a classroom [11, 13].

### 3 Study Protocol

The study was designed to investigate the relationship between students and the technology through their collaborative behaviours, and that between the teacher and the technology through her feedback and lesson plans. It was conducted over 4 sessions across a half term (6 weeks) in a UK secondary school classroom. The classroom was equipped with 8 smart tables, allowing 8 groups of 3–4 students to participate – 30 students in total. The students were native English speakers of mixed ability, studying English in Year 8 (aged 13–14, key stage 3). Each lesson was planned and facilitated by the class’s usual English teacher, who also produced the session content. Sessions were scheduled to fit in with the existing timetable. The teacher met with the research team to discuss and improve the design before the study and had completed a collaborative writing task using CCW. She designed lesson plans to incorporate the technology into her teaching goals. 2–3 Researchers were present at each session, and sessions were filmed with a classroom camera and a single group camera (Fig. 2). Before each CCW session, the students completed a collaborative exercise, either Digital Mysteries [10] (first 3 sessions) or a classroom debate (final session). The “in the wild” context also had significant practical ramifications, including:



**Fig. 2.** (a) Classroom camera (b) Single group camera

- The classroom was in use for other lessons during the day, meaning the experimental setup (i.e. all tables and recording equipment) had to be deployed before each session and dismantled after, leaving the classroom in its previous configuration.
- Schedule restrictions meant that this had to be completed in less than 1 h.
- The space available in the classroom did not allow a camera per table.

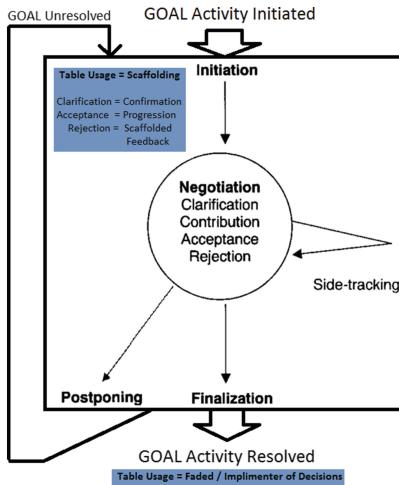
## 4 Data and Analysis

The deployment gathered data from several sources, providing multiple lenses through which to view the deployment using a mixed-methods [5] approach. A classroom camera (Fig. 2a) captured each session, the digital tabletops recorded interaction logs, the teacher recorded lesson plans and reflections for each session, and collaborative document plans and individual written work were generated and assessed by the teacher. A detailed overview of learning interactions can be derived from an analysis of the single group video (Fig. 2b) and audio-recording of each session ( $n = 4$ ). When synchronised with interaction log data (i.e. creation, manipulation and deletion of visuospatial elements, decisions made and text generated etc.), we were able to provide a detailed view of how learning was scaffolded by CCW.

This paper provides an in-depth, learning-focused analysis of a single group interacting with CCW over four sessions. Using the group as a unit of analysis [1, 28], we tracked development over time as part of scheme of work planned by the teacher. The situated nature of the deployment (i.e. in real classroom, with space and time constraints) restricted data collection opportunities (i.e. video per table) for detailed multiple group analysis. However, other data streams, such as interaction logs, classroom camera and teacher plans and reflections, provided a contextual background.

In order to be able to map a group's learning interactions to key design features of CCW (externalisation and communication of thinking through visuospatial representations) we utilised an “event-driven” approach to analysis [26]. This allowed us to evaluate, according to distributed cognition, how cognitive processes promoted by CCW were co-ordinated across time, and where events in one session impacted or transformed in later sessions [9].

Bartu [2] uses the concept of Proposals to examine group decision-making processes. Proposals fall within a class of speech acts (that can be non-verbal) which are used to make ‘suggestions’, in order to encourage listener(s) to carry out some future course of action [31]. They are performed according to how the speaker designs an action (e.g. as a question, exclamation, or imperative) [19]. The concept of Proposals maps to the design of CCW through decision points (and visuospatial representations). Based on Bartu’s notion of Proposals we developed an organisational framework to describe the decision-making process incorporating CCW (see Fig. 3). A decision-making episode is initiated through a Proposal. Then a number of different combinations of interactional moves are possible through decision-making ‘speech acts’. These moves, in conjunction with interactions with CCW, allow the speakers to implement a decision scaffolded by CCW. Analysis concentrated on the multimodal



**Fig. 3.** Expanded Bartu decision making framework

interactions at a micro-level, including verbal and non-verbal behaviour that formed the structure of the collaboration, focused on the co-construction of Decisions and use of Proposals to initiate these decision-making processes (i.e. CCW decision points).

Methods from Discourse Analysis (DA) [32] were used to facilitate a systematic and structured way to describe and analyse the organisation of group interactions. This is carried out from the perspective of distribution, co-ordination and impact of cognitive processes over time and the consequences of the design goal of collaborative decision making, i.e. “explore the organisation of social interaction through discourse as coherent in organisation and content and enable people to construct meaning in social contexts” [4]. As such, our analysis examined language and sense-making practices as they were co-constructed across multiple modes of communication including speech, gesture and other contextual phenomena [30]. The DA comprised of two phases. First, decision-making episodes were identified from the video data (independently by two researchers, followed by a consensus). These were marked as stretches of interaction made up of the initiation of a Proposal (as suggested by Barron [1]), subsequent discussion around this Proposal that ended with a decision (see Fig. 2). Secondly, in each episode, the focus of the analysis was on the incremental process of reaching a decision and implementing it (scaffolded by CCW) using Proposals. A double coding of interactional behaviours was performed; two researchers worked first independently, then together to reach agreement about the coding of decision-making ‘speech acts’ within the episode.

## 5 Results

This section provides a summary of results based on the qualitative and quantitative analysis across 3 data sets: single group video and audio, teacher notes and interaction logs. Observation of decision-making (as a product of interactions with CCW and

group members) provide quantitative data concerning use of turn-taking and Proposals. These results were used sequentially – they provided a starting point for looking in more detail at the nature and quality of the decision-making as a collaborative process mediated by CCW. Table 1 displays the number of Proposals identified in each session. Proposals are categorised and quantified (by facilitator, CCW and group), and turn taking is also quantified showing the number of speech acts overall.

**Table 1.** Proposals and turns by videoed group per session

Discourse actions	Session number			
	1	2	3	4
Number of turns by students	233	125	292	260
Proposals from facilitator/s	9	18	10	19
Proposals from students	24	26	80	61
CCW proposals (visuospatial representations, decision points)	8	19	44	53

When turns are used as a way of evaluating participation, the data suggests that Session 2 produced the lowest rate of participation in the tabletop activity. The highest number of turns, and increased participation was observed in Session 3, slightly decreasing in Sessions 4 and 5. This first phase led the initiation of specific questions concerning the relationship between turns of talk (indicating levels of participation), Proposals (i.e. those occurring away from the table, mediated directly with CCW and from facilitators), and the nature and quality of activity in the interactional space between the initiation of a Proposal and decision-making.

The following sections provide the results for the qualitative analysis of the group interactions with CCW, together with results from the analysis of the teacher's notes and interaction logs. These data were correlated according to a Convergent Parallel Mixed Methods Design [5]. For each session, an overview of the focus of the lesson is presented, followed by an excerpt providing a detailed micro-view of the nature and quality of interactions with CCW, typical of that particular stage of development of the group in the whole-class deployment. A short commentary of the transcribed episode is then provided which serves as a qualitative account.

## 5.1 Session 1 – Midsummer Night's Dream

The students worked on writing a persuasive document answering the question “Which character is the most powerful?” (~20 min) based on a Digital Mystery they had completed in the same session (~45 min). The teacher introduced the session, with topic related notes on the whiteboard, e.g. main characters, groups of characters and brief plot synopsis. The episode comes from the first ‘reading slips’ stage (Table 2).

**Commentary:** Despite the high number of turns (Table 1), suggesting a high level of participation, the nature of the interaction in terms of decision making processes (i.e. Proposals) suggests low quality. The episode shows a lack of focus on task demands,

**Table 2.** Episode from session 1: Midsummer Night's Dream

M1	right Oberon [((tries to drag and create slip but nothing is selected))]	I & P1
G	[((all watch as M1 tries to create a slip))]	C
F2	((selects a slip)) that's not the right one put it in the trash. ((moves slip to trash))	R & P 2
G	[((all watch as F1 selects and moves slip to trash))]	C & A2
M1	[in the trash]	C
F2	((continues to select slips and trash them, all continue to watch this activity))	C
M2	((selects slip and moves towards trash)) which would you like	C & P3
F2	that one	A3
M2	[((moves selected slip to the centre of the table))]	C
Fac	I((Fac moves to the group)) ... because you guys just spent half an hour reading these slides you don't need to read them all again so I'll move you onto the next stage. ((closes this part of the task and moves the group to the next stage)) now we're going to think about what paragraphs you want if you look at the instructions it will tell you how...	P4
G	((read the instructions on screen))	A4

Participants are: M1 = Male 1, M2 = Male 2, F1 = Female 1, F2 = Female 2, T = Table, G = Group and Fac = Facilitator. [] = Overlapping speech and (( )) Describes non-verbal activity. Encoding is P = Proposal, C = Contribution A = Acceptance, R = Rejection, Cl = Clarification, PP = Postponing, I = Initiation of Activity and F = Finalisation.

The episode begins with a Proposal (1) from M1 who initiates joint attention to a slip. F2 then rejects this Proposal, leading to a series of contributing turns, where F2's Proposal (2) to 'trash' existing slips is accepted by the group, until M2's Proposal (3), which is accepted, and activity continues until the Facilitator joins the group to make Proposal (4). This is accepted by each group member. There are 4 Proposals made around and through CCW, but there is no Finalisation to Proposal (1) (the initiation of the activity). Rather, the Facilitator refocuses attention on the task requirements.

evidenced by the intervention by the facilitator and in the quality of the collaboration. Although the students actively watch each other and co-ordinate their efforts, this is not integrated with CCW. Proposals and decisions are made by individuals and are superficial. Those linked to CCW are initiated without prior discussion.

Despite the lack of Finalisation, this episode demonstrates how the nature of collaboration with and around the table is linked to a mutual orientation to the task facilitated by CCW; students' joint attention to each other's talk and actions is highly co-ordinated. Students make use of the affordances of the table: demonstrating an understanding that that slips can be selected, read and, if required, trashed.

**Classroom Level:** The interaction logs of other tables in the classroom show that no groups completed the task, mostly stopping at the paragraph construction stage. Generally, task-provided evidence was seldom used with groups writing their own outline items. During the session, the teacher moved between the groups and observed

that paragraphs were being named abstractly rather than on-topic. She reflected on this in her notes, and decided to incorporate an explanation in the plan for the next session.

## 5.2 Session 2 – Midsummer Night’s Dream – Part 2

The session, exploring the same question, took around 25 min. The teacher again began with a topic summary, but also included an explanation of good paragraph naming strategies (i.e. on-topic). The teacher reminded the class of this during the paragraph creation phase. Two analysed episodes are presented, connected by the Initiation of Activity (Proposal 1) in Part 1 and the Finalisation linked to that initiation at the end of Part 2 (Tables 3 and 4).

**Table 3.** Episode from Session 2: Midsummer Night’s Dream: Part 1

M1	next stage ((uses interface to bring up options for move to next stage))	I & P1
T	((displays next stage confirm))	C
G	((all confirm next stage))	A1
T	[((displays message to indicate that the activity cannot progress as not enough paragraphs have been selected by the group))	R/I/P2

*In Part 1, M1 initiates the main activity by proposing a move to the next stage in CCW (Paragraph Creation). After Clarification from CCW, all display an Acceptance of M1’s Proposal, indicating a consensus. CCW warns that there are not enough connections (Proposal 2). In Part 2, the decision-making process is expanded to take into consideration this application-based Initiation of Activity.*

**Commentary:** Table 1 suggests a lower rate of participation in terms of turns, however there was a relative increase in Proposals (from facilitators, students and CCW), suggesting a closer task focus from the students and the teacher.

The episodes, as representative of broad patterns of interaction in Session 2, demonstrate an improvement in the quality of collaboration from the first session. Co-ordinated multimodal interaction shows behaviours focussed on specific goals, scaffolded more specifically here by CCW. However, similar Session 1, there is little verbalisation, so actions on the table are not considered by the group before being executed. There is also a continuation of similarly individualised action when M1 directs the connecting of paragraphs to F1. Proposals are generated by students and CCW, co-ordinated to achieve specific interactional goals. CCW and the teacher play a more concrete role in scaffolding decision-making by generating Proposals and mediating those put forward by the group.

**Classroom Level:** The logs show that groups changed their paragraph naming strategy after the mid-session reminder and all groups finished the task although some rushed rather than create good answers (particularly with regard to connections and use of evidence). Again, groups did not use the provided evidence, preferring to write their own. Rather than use the generated plans for a text generation exercise, the teacher created a model answer for the students to look at in the next session.

**Table 4.** Episode from Session 2: Midsummer Night's Dream: Part 2

G	[((all read the information displayed))	C
M1	[more paragraphs oh then you need to do it	C/P2
G	[((all confirm that they have read the message and move back to the on-going activity of connecting paragraphs))	C
F1	((selects previously written text which has not been connected and drags to paragraph))	A2
T	((displays the two paragraphs which the students wish to link and connection dialogue))	P3
M1	((selects connective))	C
T	((connective is added to the displayed text))	C
G	((all re-read the text with the new connective in place))	C
F1	most powerful is IS ((shows group where she is referring to))	C/P4
F2	((modifies the text in the paragraph))	A4
G	((confirm that they agree to all the changes))	A3
F1	((selects next stage))	A1/F/I

*Part 2 opens with an all-group Contribution, which is built on by M1, via a verbalised contribution and an additional Proposal (2) addressed to F1. This initiates her to connect the last paragraph. Proposal (2) overlaps with an additional all-group Contribution where all students confirm that they have understood the need to continue connecting. Then, F1's Acceptance displays her joint attention to M1 and Proposal (2) offered by CCW when she connects two paragraphs. CCW offers another Proposal (3) – the paragraph creation dialog. Individual and collective Contributions then follow until Proposal (4) by F1 who initiates specific joint attention to an error. F2, who is holding the keyboard, makes the correction, which is accepted by F2 and the group. F1 then proposes a move to the next stage, the original activity initiated in Proposal (1) and the students move to the next stage – a Finalisation point.*

### 5.3 Session 3 – Greek Mythology

Before the session, the teacher asked students to “assess” her previous session answer. In this session (~25 min), the aim was to create a persuasive document about Greek Mythology. The teacher provided a summary on the board, and the context – writing a Proposal for a museum display. Three episodes are analysed showing how the organisation of the decision-making processes around CCW (based on Proposals) are not isolated events but evolve incrementally, intertwined with previous decisions (Tables 5, 6 and 7).

**Commentary:** The session had the highest rate of participation in terms of turns, and also the highest number of Proposals, both from the students and CCW. However, the Proposals provided by the facilitator decreased. The nature and quality of the decision making process increased, as demonstrated in the representative episode.

Students continue to demonstrate joint attention to the task as they watch the text appearing on the table. Central to this change is the occurrence of talk prior to finalisation of decision-making which is implemented using the table. The focus on the content of the paragraphs shows how the Proposals are used in the development of finalisation rather than to initiate it – indicating a change in the distribution of who is

**Table 5.** Episode from Session 3: Greek Mythology: Part 1

F2	right monsters ((holds keyboard and prepare to type))	I/P1
M1	no gods	R1/P2
M2	just put gods	C/R1
F2	goddesses or gods	R1/P3
M1	Gods	C
M2	gods and then goddesses.	C
F2	((types “gods” as Paragraph title))	A2
G	[(all watch the interface as F2 types in the text))	C
M1	[just put gods and goddesses.]	C/R1/P 4
F2	[(confirms Paragraph Creation)) no it’ll work better this way, because it means then we’ll have more excuse to do more writing.	R3/A2/F

*Part 1 opens with a Proposal (1) from F2, who is holding the keyboard and acting as ‘scribe’ to initiate the activity of creating a paragraph. The title of the paragraph (rather than the creation) is met with a Rejection from M1 who proposes an alternative (Proposal (2)). In subsequent turns of this decision-making process, there are a number of contributions (6) and Rejections (3) until Proposal (2) is accepted by F2. A further Proposal (3) is also rejected in this same turn. This complex sequence (which does not reach complete Finalisation) shows specific joint attention to the ‘content’ of Proposal (1): There is a collaborative orientation to correct terminology where it is shown to be an explicit part of the decision-making process.*

**Table 6.** Episode from Session 3: Greek Mythology: Part 2

F1	((Selects Creates Paragraph))	P1
F2	((types “Goddesses” as Paragraph title))	C
G	(all watch the interface as F2 types in the text))	C
M2	yes, it’s actually two Ds you can just use the arrow keys.	PP/P2
M1	I’m sure it’s not two Ds.	C/R2
F2	It’s not two Ds. ...discussion as to the spelling of goddesses continues over a number of turns...until group creates new paragraph	C/R2 A1/F

*Part 2 shows similar decision-making processes based around the content of paragraph headings. This time, the spelling of the title is proposed (Proposal (2)) for joint consideration by M2 which is subsequently rejected by others within the group until the paragraph is created.*

leading the process and the role of CCW. These episodes demonstrate a change in terms of the development of decision-making processes from Sessions 1 and 2.

**Classroom Level:** Logs show that all groups continued in the “on topic” paragraph naming strategy, but created more topics in less time. All groups created plans that were assessed by the teacher. The teacher reported that writing a “proposal”, rather than a “straight forward” persuasive document, was “too much” for some groups, and decided to go over the “basics” of persuasive writing and choose an easier context.

**Table 7.** Episode from Session 3: Greek Mythology: Part 3

F1	((selects creation of new Paragraph))	P1
F2	right what else	C
M1	demi-gods	C
F2	((types demi-gods as Paragraph title))	C
G	((all watch the interface as F2 types in the text))	C
M1	shall we put demi-gods and demi-goddesses in the same thing	CL/P2
F2	yes Courtiers wasn't it no that's in A Midsummer Night's Dream ((confirms Paragraph Creation))	CL/R2/A1
F1	((Selects Creates Paragraph))	F

*Part 3 shows how, in the initiation of a third activity to create a paragraph with a title, there is joint attention to the activity shown through the number of individual and collective (verbal and non-verbal) Contributions, where M1's offering is taken up. As F2 types, M1's joint attention leads to a further Proposal (2) to expand the title. Negotiation involves F2, using both verbal and non-verbal means to reject this Proposal by completing her typing and confirming the paragraph creation, leading to finalisation.*

#### 5.4 Session 4 – Sport vs. Library

In this session the students were working on writing a persuasive argument to decide between funding for a library or new sports facilities at the school (~25 min). They had previously held a classroom debate, including a paper based exercise involving reading and organising evidence and producing a structure they could use to design their document. (Interestingly, the structures mirrored the CCW process without prompting). The teacher provided a short reminder of persuasive texts (Table 8).

**Commentary:** Table 1 indicates a high rate of participation in terms of turns (compared to first and second sessions). The number of Proposals remains similar to session 3, however CCW is being used more to implement and scaffold decisions.

As one of a number of similar episodes observed in Session 4, this episode shows how the collaborative nature of decision-making processes with and around CCW has developed across the sessions; culminating in co-ordinated, productive and high quality multimodal interaction. The students actively watch and listen to each other and collaboration contains both individualised and collective Proposals. CCW is used to both initiate Proposals by members of the group (dragging and dropping slips into the paragraph) and support the development of verbal Proposals linked to content which is discussed prior to being added. High quality collaboration is not defined by one decision-making structure but a merging of events built up over the sessions.

**Classroom Level:** All groups completed the task, creating plans that were used as the basis for an individual writing homework exercise, assessed by the teacher. The teacher was encouraged by the final documents, reporting that all students showed improvements.

**Table 8.** Episode from Session 4: Sport vs Library

F2	for the library you could have ((reads notes she has on lap))	P1
F1&M1	((look at a distance at the page of notes that F2 is reading))	C
F2	((Selects Slip and drags to Library Paragraph))	P1
F1	((Selects another Slip and drags to Library Paragraph))	P2/A2
F2	I knew all of them just put like number three put that in then write something.	C/P3
M1	((begins to type into the library paragraph))	A3
M2	like where we can allow children to study hard for upcoming [exams	P4
M1	[expand their learning [skills	P5
M2	[expand their knowledge their knowledge on	P6
G	((all watch as M1 adds the additional text to the paragraph))	A4/5/6
M1	on core subjects	P7
G	((all nod their heads in agreement))	A7
F2	special subjects	P8
M1	yes [(types the points into the library paragraph))	A7/8
M2	[then put like i.e. English, Maths.	P9
M1	don't know if they need special case ((M1 ((types final point into paragraph)))	CL/R/F
G	[(watch M1 as he types final point))	C
M1	What could be used against the library?	I/P1

The episode opens with an Initiation of Activity by F2 to add content to a paragraph 'For Library'. In doing so, the group's joint attention is on F2 as she consults her notes then drags a slip into the Library Paragraph. This Proposal (1) is followed by a similar action by F1 who drags another slip to the same box. In F1's turn there is no dialogue and no other group member accepts nor rejects this Proposal. F2 next proposes a new action (Proposal (3)) to write something in the Library Paragraph (rather than only drag existing slips). Eight overlapping Proposals then follow and are accepted by 1) M1 who is the 'scribe' on this occasion and notes them down via the keyboard 2) non-verbal means such as nodding and 3) echoing and building on each other's turns (Proposals). The activity ends when a group display of joint attention to M1 who types the final point into the paragraph, which leads to the initiation of a new, linked activity ("what could be used against the library?").

## 6 Discussion

In this study, we used a single group analysis to obtain a deep understanding of activity at the tabletop throughout a classroom deployment, but also used further classroom data streams to provide a whole class context. We were able to integrate the technology into the analysis of the group's discourse due to specific design elements of CCW designed to generate Proposals (i.e. decision points), and by adapting Bartu's model (Fig. 3) to recognise these Proposals alongside those generated through discourse. The multi-tabletop classroom deployment has enabled evaluation of CCW as a learning tool (evidenced through the students' changing relationship with CCW), as well as provided insights into the teacher's integral role in integrating technology in the classroom (evidenced through the teacher's reflection and adaptation of lesson plans).

## 6.1 Student's Relationship with CCW

Students had a changing relationship with CCW, which centred on Proposals, and their talk developed across the sessions. The group level analysis allowed us to view the different kinds of collaborative events driving the learning in specific interactions around decision-making processes over time in terms of student, facilitators and CCW. Participants' use of specific design elements, included to elicit collaborative behaviour such as decision points and visuospatial elements, were also developed.

In the first two sessions, facilitators were the prime source of Proposals and CCW was not central to the students' talk nor to their activity. CCW Proposals (i.e. through decision points) were used superficially, in an individualised manner and with little follow-on development. In these initial sessions, paragraphs were created with abstract names (until the teacher's classroom intervention) and task-provided evidence was seldom used. Rather, students made their own points for the paragraphs.

In Session 3, group Proposals were offered and discussed while CCW was used to facilitate the transition of ideas to the plan as a collaborative effort. CCW provided scaffolding via Proposals (decision points etc.) that were attended to by the students. Eventually, CCW could be seen to begin to 'fade' [35], with more talk happening off-table before using CCW (to confirm "correctness" before proceeding). Paragraphs were created with on-topic themes, although paragraph connection was still naive. Evidence, was now being used extensively and correctly in persuasive arguments.

In Session 4, Proposals largely came from the group, while CCW mainly *implemented* shared decision making, i.e. interactional focus was more on the students than the table. Students decided what they wanted to do via talk *before* implementing their decision on the table. In fact, some CCW design elements (i.e. prompting for group agreement) began to hinder students' collaboration. The scaffolding provided by CCW "faded", but the application did not recognise this and still provided the same mechanisms. Task-provided evidence was used in conjunction with the students' points rather than an all or nothing strategy from previous sessions, i.e. to corroborate persuasive arguments being generated by the students.

As this study has evidenced, the identified changing relationship between students and CCW has clear generalizable consequences on application design. They necessitate that the application design should be flexible enough to allow for such variance in usage. The application should therefore allow for controlling (whether automatically or even manually by the teacher or students) the level of scaffolding, and correspondingly fading, provided. Otherwise, the application's role may shift from being supportive of the task to being a burden. This variable scaffolding could also be a tool for differentiation across groups of differing abilities across the classroom.

## 6.2 Integration into the Classroom

A key factor for the impact of the study is how the technology was integrated into the classroom. The teacher is vital in this process, as they control the teaching goals and orchestrate the classroom based on their familiarity with the needs of the students.

Looking at the additional data streams (classroom camera, interaction logs and the teacher's lesson plans and notes), we can monitor the bigger picture of the classroom

over the study. They show how the teacher adapted her plans, during or after the sessions, to get the best out of the technology. She used classroom introductions and interventions to provide background information and suggested strategies for successful persuasive writing (such as on-topic paragraphs). She adapted her assessment approach when it became clear that the students would not complete the task (by supplying a model answer for students to “mark” rather than writing their own). She also appropriated the sessions for use beyond the persuasive writing task, i.e. as a method for bringing out the underlying theme of “power” from the students.

As a class, the data show that the students benefited from the teacher’s adaptive, student-centred teaching approach. The work produced, such as creating and connecting paragraphs and using evidence, increased in quality, but not always linearly. E.g. in Session 3, some groups had lost a focus on persuasive writing, and relied on existing evidence provided by the teacher via CCW. Session 4, saw a move to more persuasive-orientated, student-generated content. In summary, the technology gave the teacher more power to fulfil her teaching goals and benefit the class, rather than replacing some or all of her activities.

The in-the wild nature of the study places a premise on the data collection process. Time and space constraints do not allow for every group to be individually recorded, and so detailed findings from a single group must be combined with higher level data from the classroom level. Examination of interaction logs, classroom camera and teacher notes and reflection allow for general patterns to be identified that indicate that the class as a whole followed a similar pattern to the single group. However, in an ideal world, full data on all groups would yield a more comprehensive analysis.

Despite such limitations, these observations were only possible due to our “in the wild”, classroom based approach. Two important design elements of the application allowed the teacher to make the required observations to adapt her strategy during the sessions (and across the study as a whole):

1. The externalization concepts incorporated into the design and cumulative representation aspects of the design raised the teacher’s awareness of the students’ tendency to use superficial paragraph names and lack of persuasive line of argument, rather than waiting until a post-session outcome assessment.
2. The flexibility of the design enabled the teacher to implement these changes in strategy in the introduction or mid-session classroom announcements, such as assessing group-made plans before individual writing assessment, providing a sample document for students to consider, or focusing on particular task demand.

## 7 Conclusion and Future Work

We presented an “in the wild” evaluation of a Collocated Collaborative Writing application on multiple digital tabletops. We used video data integrated with interaction logs to gain a detailed moment to moment insight into one group’s learning. We used a classroom camera, interaction logs and teacher plans and reflection to build a context-specific analysis.

We observed that the students' collaboration progressed from relying on teacher and facilitator Proposals, through using CCW's scaffolding, to the point where CCW became simply a planning and assisting tool where the scaffolding aspect faded. We found that the visibility of the state, made possible through the cumulative nature of the task, along with the flexibility of dividing the task into stages, enabled the teacher to adapt her lesson plans over the course of the deployment. Moreover, the role of the teacher as classroom orchestrator in conjunction with the technology was vital, requiring a positive relationship with the technology with regard to her teaching goals.

This study has worked towards identifying and understanding more about the nature and quality of learning through CCW in an 'in the wild' classroom deployment. It discussed the deployment of the application as part of the 'machinery' of everyday classroom life. In doing so, the study's mixed method approach allowed us to build up an understanding of how the relationship between the students and technology, as well as the teacher and technology, changed over time in the everyday life of the classroom. Some of these behaviours and decisions may prompt for new approaches to teaching and learning (for example, new ways of looking at group collaboration alongside individual work). Yet they also demonstrate how the teacher can still maintain existing practices in terms of classroom orchestration, i.e. technologies can be normalised into the classroom.

This change in relationship between students and technology, which was only observable because of the longitudinal, "in the wild" nature of our study, is not specific to CCW. It demonstrates that any collaborative learning technology targeting the classroom should be flexible enough to take account of such change in relationship, and make such changes clearly visible to the teacher. The design should provide visibility of both task-level details (paragraph names for example), and also higher level usage patterns (such as the overall level of interaction with the technology and communication amongst the group). By allowing such visibility of use, combined with flexibility in implementing teacher's changes in strategy (whether 'on the fly' in-session, or between sessions), it is possible to have the positive relationship with the technology, as shown in this study, which is key to technology adoption in the classroom. This would not have been possible in a lab-based or single-group deployment.

This visible, flexible teacher orientated design approach raises possibilities for further investigation into how such flexibility can be further supported by applications - not only how to support (and design for) different ability levels, but make scaffolding dynamic or adaptive as this relationship changes. This initiated classroom-sensitive questions such as: How much influence should the teacher have over these adaptations? Are they part of the initial lesson plan, or can they be tweaked on the fly? Can orchestration tools be developed that allow teachers to monitor and adapt tasks to enable dynamic differentiation?

## References

1. Barron, B.: When smart groups fail. *J. Learn. Sci.* **12**(3), 307–359 (2003)
2. Bartu, H.: Decisions and decision making in the Istanbul exploratory practice experience. *Lang. Teach. Res.* **7**(2), 181–200 (2003)

3. Berninger, V., et al.: Assessment of planning, translating, and revising in junior high writers. *J. Sch. Psychol.* **34**(1), 23–52 (1996)
4. Coyle, A.: Discourse analysis. In: Breakwell, G.M., Hammond, S., Fife-Schaw, C. (eds.) *Research Methods in Psychology*. Sage, London (1995)
5. Creswell, J., Clark, V.P.: *Designing and Conducting Mixed Methods Research*. Sage Publications, Thousand Oaks (2010)
6. Dillenbourg, P.: Distributing cognition over brains and machines. In: Vosniadou, S., De Corte, E., Glaser, B., Mandl, H. (eds.) *International Perspectives on the Psychological Technology-Based Learning Environments*, pp. 165–184. Lawrence Erlbaum, Mahwah (1996)
7. Dillenbourg, P., Evans, M.: Interactive tabletops in education. *Int. J. Comput. Collab. Learn.* **6**(4), 491–514 (2011)
8. Heslop, P., et al.: Learning extended writing: designing for children's collaboration. In: *Proceedings of 12th International Conference on Interaction Design and Children*, pp. 36–45 (2013)
9. Hollan, J., et al.: Distributed cognition: toward a new foundation for human-computer interaction research. *ACM Trans. Comput. Interact.* **7**(2), 174–196 (2000)
10. Kharrufa, A., et al.: Digital mysteries: designing for learning at the tabletop. In: *ACM International Conference Interaction Tabletops Surfaces*, pp. 197–206 (2010)
11. Kharrufa, A., et al.: Extending tabletop application design to the classroom. In: *Proceedings 2013 ACM International Conference Interaction Tabletops Surfaces - ITS 2013*, pp. 115–124 (2013)
12. Kharrufa, A., et al.: Learning through reflection at the tabletop: a case study with digital mysteries. In: *World Conference Education Multimedia, Hypermedia Telecommunication*, pp. 665–674 (2010)
13. Kharrufa, A., et al.: Tables in the wild: lessons learned from a large-scale multi-tabletop deployment. In: *CHI* (2013)
14. Kharrufa, A.S., Olivier, P.: Exploring the requirements of tabletop interfaces for education. *Int. J. Learn. Technol.* **5**(1), 42 (2010)
15. Kirkpatrick, L.C., Klein, P.D.: Planning text structure as a way to improve students' writing from sources in the compare-contrast genre. *Learn. Instr.* **19**(4), 309–321 (2009)
16. Kirsh, D.: The intelligent use of space. *Artif. Intell.* **73**(1), 31–68 (1995)
17. De La Paz, S., Graham, S.: Explicitly teaching strategies, skills, and knowledge: writing instruction in middle school classrooms. *J. Educ. Psychol.* **94**(4), 687–698 (2002)
18. Lewis, M., Wray, D.: *Writing Frames. Reading and Language Information Centre*, University of Reading, Reading (1996)
19. Markee, N.: *Conversation Analysis*. Lawrence Erlbaum, Mahwah (2000)
20. Martinez-Maldonado, R. et al.: Orchestrating a multi-tabletop classroom: from activity design to enactment and reflection. In: *Proceedings of 2012 ACM International Conference on Interaction Tabletops Surfaces ITS*, pp. 119–128 (2012)
21. Mercier, E.M., Higgins, S.E.: Collaborative learning with multi-touch technology: developing adaptive expertise. *Learn. Instr.* **25**, 13–23 (2013)
22. Norman, D.A.: *Things that Make Us Smart*. Perseus Books, Boston (1993)
23. Piaget, J.: *The Language and Thought of the Child*. Routledge, New York (2002)
24. Rick, J., et al.: Beyond one-size-fits-all: how interactive tabletops support collaborative learning. In: *Proceedings of the 10th International Conference on Interaction Design and Children*, pp. 109–117. ACM (2011)
25. Rick, J., Rogers, Y.: From DigiQuilt to DigiTile: adapting educational technology to a multi-touch table. In: *TABLETOP*, pp. 79–86. IEEE, Los Alamitos (2008)

26. Rogers, Y.: HCI Theory: classical, modern and contemporary. *Synth. Lect. Hum.-Centered Inform.* **5**(2), 1–129 (2012)
27. Rogers, Y., Ellis, J.: Distributed cognition: an alternative framework for analysing and explaining collaborative working. *J. Inf. Technol.* **9**(2), 119–128 (1994)
28. Roschelle, J., Teasley, S.: The construction of shared knowledge in collaborative problem solving. In: O’Malley, C. (ed.) *Computer Supported Collaborative Learning*. NATO ASI Series, vol. 128, pp. 69–97. Springer, Heidelberg (1995)
29. Schneider, B., et al.: Phylo-Genie: engaging students in collaborative ‘tree-thinking’ through tabletop techniques. In: CHI 2012, pp. 3071–3080 (2012)
30. Scollon, R., Levine, P.: Multimodal discourse analysis as the confluence of discourse and technology. In: Scollon, R., Levine, P. (eds.) *Discourse Technology Multimodal discourse Analysts*, pp. 1–6. Georgetown University Press, Washington, DC (2004)
31. Searle, J.R.: *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, Cambridge (1969)
32. Sinclair, J.M., Coulthard, M.: *Towards an Analysis of Discourse: The English Used by Teachers and Pupils*. Oxford University Press, London (1975)
33. Vygotsky, L.S.: *Mind in Society*. Harvard University Press, Cambridge (1978)
34. Wood, D., et al.: The role of tutoring in problem solving. *J. Child Psychol. Psychiatry* **17**(2), 89–100 (1976)
35. Wood, D., Wood, H.: Vygotsky, tutoring and learning. *Oxf. Rev. Educ.* **22**(1), 5–16 (1996)
36. Von Zadow, U., et al.: SimMed: combining simulation and interactive tabletops for medical education. In: CHI 2013, pp. 1469–1478 (2013)
37. Zhang, J., Patel, V.L.: Distributed cognition, representation, and affordance. *Pragmat. Cogn.* **14**(2), 333–341 (2006)

# Evaluating the Accuracy of Pre-kindergarten Children Multi-touch Interaction

Vicente Nacher<sup>(✉)</sup> and Javier Jaen

ISSI Group, DSIC, Universitat Politècnica de Valencia, Valencia, Spain  
vnacher@dsic.upv.es, fjaen@upv.es

**Abstract.** The direct manipulation interaction style of multi-touch technology makes it ideal for pre-kindergarten children. Recent studies have shown that these challenging users are able to perform a set of basic multi-touch gestures. However, little is known about the accuracy that they can achieve. This paper evaluates the performance of pre-kindergarten children when accuracy is required in the termination phase of these gestures and points out that a mechanism for dynamically adapting the accuracy level could help children in their motor skills development.

**Keywords:** Multi-touch interaction · Gestures · Usability evaluation · Pre-kindergarten · Accuracy

## 1 Introduction

Nowadays children between zero and eight years old are frequent users of digital media [3]. In fact, as touch allows a more intuitive and natural way of interaction [7], they are often exposed to multi-touch technology even before they learn higher oral communication skills.

This has been confirmed by recent works such as [6] which reveals that even children between the ages of two and four are able to perform a basic set of touch gestures and [9] which concludes that, overall, children aged 3 to 6 years are able to perform the tap, double tap, drag & drop and double drag & drop gestures. However, these works have also pointed out that very young children have precision problems in both the acquisition and termination phases of these interactions because of limitations in cognitive and motor skills.

Regarding this matter, there are no studies in the literature addressing the topic of accurate performance of multi-touch gestures by pre-kindergarten children. In this paper we explore whether children aged two to three years are able to perform a set of touch gestures when high levels of precision are required and evaluate whether factors such as age and gender have an impact on the performance of these interactions by pre-kindergarten users. This paper contributes to a growing body of literature in the area of children-computer interaction by providing findings from a controlled experiment with four touch gestures in which the termination phase must be performed with high-levels of precision. The findings will confirm that, at this early age-range, there are significant differences among subjects with respect to precision and, therefore, designers of future touch based applications for these specific users should devise

adaptive mechanisms to cope with different levels of accuracy that allow very young children to exercise and incrementally develop fine-grained touch interaction skills.

## 2 Related Works

Several studies have analyzed the use of touch devices by pre-kindergarten children. The works of Abdul Aziz et al. [1, 2] evaluated the tap, drag, rotate, drag and drop, pinch, spread and flick gestures with children aged 2 to 4 years. Their results showed that 4 years old children were able to perform all gestures, the 3 years old ones only had some issues with the spread task and the youngest users (2 years old) were able to perform the tap and drag gestures properly but had some issues with the more complex ones.

On the other hand, the study of Nacher et al. [6] evaluated a basic set of multi-touch gestures with pre-kindergarten children (2 to 3 years old) and concluded that they are able to perform gestures such as tap, drag, scale (up & down) and one finger rotation. Moreover, the authors of this work point out that, when some proposed assisted strategies are used, pre-kindergarten children are able to perform more problematic gestures such as double tap and long press [5] with high success rates.

Vatavu et al. [9] evaluated touch gestures (tap, double tap, drag & drop, multiple drag & drop) with children aged between 3 and 6 years on tablets and smartphones. Their results showed that although all children had high success rates, there was a significant performance increase with age in terms of success rate and time spent performing the gesture which is an expected behavior.

These works seem to conclude that pre-kindergarten children have the necessary skills to make use of multi-touch technology. However, they assume that children may not accurately perform the multi-touch gestures under consideration and always implement assistive techniques to deal with precision issues during the initiation and termination phases of each gesture. This results in interaction styles in which pre-kindergarteners do not have the control over the termination of the gestures despite some of them have the proper cognitive abilities to perform the gestures with higher levels of precision. As a result, existing applications designed under these assumptions do not benefit from the use of multi-touch technology to help children to develop their precision-related cognitive and motor skills. Hence, in this work we evaluate the drag, scale up, scale down and rotation gestures when accurate termination of the gestures is required. The goal is to gain additional knowledge about precision issues when pre-kindergarten users are considered in the design of future touch-based applications. The results of this work would allow the design of applications that provide assistive strategies to deal with precision issues in an adaptive way only for less skilled children and not in an exhaustive way for every child as current systems do.

## 3 Experimental Study

The overall goal of our experimental study is to identify whether pre-kindergarten children are able to get high success rates when performing gestures with high accuracy levels. More specifically, the research questions of the study are formulated as follows:

*When high accuracy touch gestures are considered...*

*RQ1: ...is the degree of success independent of age group?*

*RQ2: ...is the degree of success independent of gender?*

*RQ3: ...is the completion time independent of age group?*

*RQ4: ...is the completion time independent of gender?*

*RQ5: ...is the average error independent of age group?*

*RQ6: ...is the average error independent of gender?*

### 3.1 Participants

Forty children aged between 25 and 38 months took part in the experiment (Mean ( $M$ ) = 31.60, Standard Deviation ( $SD$ ) = 4.32). Children were balanced in gender and in age group, i.e., two age groups 24 to 30 months and 31 to 38 months, with 10 males and 10 females per group were configured. Participants from two Spanish nursery schools were involved in order to explore whether children could perform non static touch gestures, i.e. requiring the movement of contacts across the surface. Parental consent was obtained before carrying out the study.

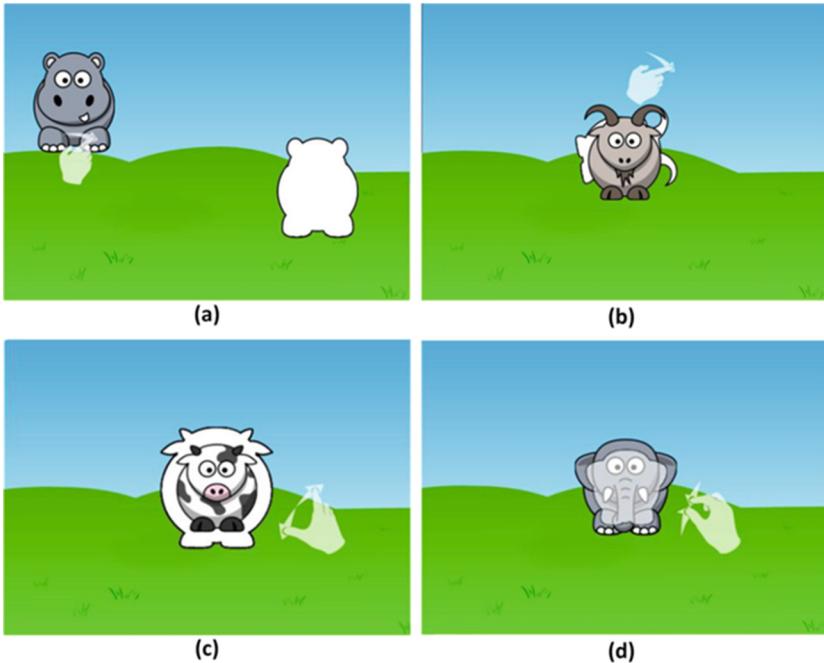
### 3.2 Apparatus

The interaction framework for the experiment was implemented in Java using JMonkeyEngine SDK v.3.0beta. The devices used for the experiment were a Motorola MZ601 and a Samsung Galaxy Note 10.1 tablet with Android 3.2 both with capacitive multi-touch screens.

### 3.3 Tasks

**Drag.** Astatic image of an animal appears in a random position on the screen and the same (reference) image appears in a white profile in another random position, always at a distance of 588 pixels so as to be able to compare execution times among the different subjects (Fig. 1-a). The random position of the reference image is subject to some geometric restrictions, to make sure that it is completely visible on the surface. Participants are requested to drag the target to the reference image with one finger. The task is successful when the target image reaches the location of the reference image with a precision of less than 15 pixels on both X and Y axis when the subject lifts his/her hand (like a drag and drop gesture).

**Rotation.** A static image of an animal appears in the center of the screen in front of a blank profile of the same image in a different orientation. Rotation is always clockwise to a fixed position so as to be able to compare interaction execution times among subjects (see Fig. 1-b). Participants are requested to rotate the target image to the position of the reference image by dragging one finger around the center of the target image. Pressure can be applied on the target image itself or anywhere around it. The



**Fig. 1.** Example tests: (a) drag (b) rotation (c) scale up (c) scale down

task is successful when the target image reaches the orientation of the reference image with a precision of less than  $5^\circ$  when the subject lifts his/her hands.

**Scale up.** A static image of an animal with a size of 5 cm appears in the center of the screen within a similar but 1.5 times larger reference shape (see Fig. 1-c). Participants are requested to scale up the target image to the size of the reference shape. This can be done by expanding the distance between two fingers of either one hand or two hands. The fingers do not have to be in contact with the reference image and the scaling factor applied is the incremental value returned by the JMonkeyEngine runtime for this gesture. If more than two contacts are made on the surface, JMonkeyEngine considers only the two most recent ones for communicating scaling events. The task is successful when the size difference between the manipulated and the reference images is less than 5 % when the subject lifts his/her hands.

**Scale down.** A static image of an animal with a size of 15 cm appears in the center of the screen superimposed on a similar reference shape half its size (see Fig. 1-d). Participants are requested to scale down the target image by making the target object shrink until it reaches the size of the reference image using two fingers of either one or two hands. The task is successful when the size difference between the manipulated and the reference images is less than 5 % when the subject lifts his/her hands.

### 3.4 Procedure

For each task, the children participated in a 2-minute learning session with an instructor. Then, the experimental platform asked them to perform the task with no external adult intervention. They had to perform five repetitions of each gesture as described in the Tasks section. When the gesture was successfully completed, the platform gave a positive audiovisual feedback. If the instructor observed that the participant did not carry out the task in a given time, it was marked as unsuccessful and the child went on to the next one. For each interaction, the system recorded the start time (seconds needed to go into action after the test began), completion time (milliseconds until the gesture was completed), success (performed correctly or incorrectly), and the number of times that users lift their hands while performing the gesture. Additional notes were taken by an external observer for posterior analysis.

## 4 Results

### 4.1 Success

In order to aggregate the success variable over the five repetitions, the variable was expressed as a percentage according to the number of repetitions performed successfully (Table 1).

The results of a two-way between-subjects ANOVA on the success revealed a significantly main effect of the *age group* factor in the scale up ( $F(1,40) = 8.052$ ,  $p = .007$ ), scale down ( $F(1,40) = 10.913$ ,  $p = .002$ ) and rotation ( $F(1,40) = 5.930$ ,  $p = .020$ ) tasks but not in the drag task ( $F(1,40) = .951$ ,  $p = .336$ ). The statistical analysis of the *gender* factor only revealed a significant effect in the rotation task ( $F(1,40) = 10.104$ ,  $p = .003$ ).

### 4.2 Completion Time

The average of each subject's successful tasks is used to obtain the completion time aggregated by users (see Table 2). The unsuccessful tests were not included in the completion time analysis.

**Table 1.** Success rate of each task by group.

Task	Gender		Age Group		Overall
Drag	<i>F</i>	<i>M</i>	<= 30	>30	78.50
	82	75	73	84	
Scale up	<i>F</i>	<i>M</i>	<= 30	>30	42
	46	38	27	57	
Scale down	<i>F</i>	<i>M</i>	<= 30	>30	45.50
	48	43	27	64	
Rotation	<i>F</i>	<i>M</i>	<= 30	>30	60
	77	43	47	73	

**Table 2.** Completion time in milliseconds of each task by group.

Task	Gender		Age Group		Overall
	F	M	<= 30	>30	
Drag			13721.63	13793.40	13755.53
			16922.11	10588.94	
Scale up	F	M	<= 30	>30	9991.62
			9517.72	10402.34	
Scale down	F	M	<= 30	>30	9859.62
			10109.92	9591.46	
Rotation	F	M	<= 30	>30	16651.61
			16972.38	16239.20	
			17751.49	15681.13	

The conducted two-way between-subjects ANOVA on the completion time revealed significant main effects of the *age group* factor for the drag task ( $F(1,36) = 7.844$ ,  $p = .009$ ) where the older group performed significantly faster. No other effects were found significant for any gesture and factor.

### 4.3 Accuracy

In order to evaluate pre-kindergarten performance when accurate termination of the gestures is required, in this section the error values for each gesture by age group and gender are showed. For each task the error is calculated as the discrepancy between the reference and the manipulated elements. In the case of the drag gesture the error is measured as a distance in pixels between them, for the scale-up and down gestures the error is a percentage measuring the discrepancy of size and for the rotation gesture the error is measured in degrees as the difference between their rotation values.

On the one hand, the results show that both age groups have similar levels of accuracy when terminating the drag gesture, i.e., similar average error ( $F(1,39) = .179$ ,  $p = .675$ ) with  $\text{avg\_error}(\text{drag, young}) = 10.75\text{px}$  and  $\text{avg\_error}(\text{drag, old}) = 8.86\text{px}$ .

On the other hand, the results show that in the scale (up & down) tasks the older group had almost a 50 % greater precision than the younger one ( $F(1,72) = 10.885$ ,  $p = .002$ ) with  $\text{avg\_error}(\text{scales, young}) = 20.84\%$  and  $\text{avg\_error}(\text{scales, old}) = 10.74\%$ . This is also the case for the rotation task in which older children have a 50 % greater precision ( $F(1,37) = 6.497$ ,  $p = .016$ ) with  $(\text{avg\_error}(\text{rot,young}) = 29.7^\circ)$  and  $\text{avg\_error}(\text{rot,young}) = 12.06^\circ$ ). Finally, with respect to the gender factor no significant differences were found for any task.

## 5 Discussion and Future Work

There are several interesting conclusions that are obtained when considering multi-touch gestures with high-levels of termination precision. Firstly, the results reveal that age is the main factor affecting the degree of success variable (RQ1) for all tasks except for the drag gesture. In this specific gesture all children, no matter what age, achieve similar high success rates ranging from 73 % to 84 %. However, for more

complex tasks such as rotations and two-finger scaling, the additional requirement of precise termination of the gesture has a high impact on less than 30 months children. Success rates decrease for this age group (27 %–47 %) and are significantly better for children aged 31 months and over (57 %–73 %). This result confirms that children start to develop their fine motor skills at this age [6] and, therefore, applications requiring higher levels of precision could be designed to stretch and challenge children in the second age group. This would be in concordance with the principles of differentiated instruction [8]. Secondly, it was also observed that girls are on average more successful than boys when precision is an issue (RQ2). These results are consistent with previous work which shows superior fine motor skills in girls [4]. However, our study only revealed significant differences for the rotation task. This is because the additional difficulty associated to the coordination of two finger contacts makes the scale-up/down tasks specially challenging for both boys and girls when precision is required (see Table 1).

In addition, if completion time is considered (RQ3 & RQ4), no significant differences were found in terms of age nor gender for all tasks except for the drag gesture. This means that relatively challenging actions in terms of cognitive and motor skills such as the scale-up/down and rotation gestures are performed by all children at similar speeds. It was observed that the additional precision requirement in these gestures forced all children, no matter their age or gender, to perform the final phase of the interaction (contact release phase) repeatedly until the final successful completion of the gesture was achieved. This was not the case for the drag gesture (see Fig. 2), specially perceived by older children (aged 31 months and over) as an easy to perform action they were able to complete with a lower number of attempts and, thus, resulting in significantly lower completion times.



**Fig. 2.** Pre-kindergarten child performing the drag task.

Finally, when accuracy is considered (RQ5), the analysis of the average errors reveals that the most challenging gestures (scale up-down and rotation) are performed with significantly higher levels of accuracy (lower average error) by older children (see

Sect. 4.3). This observation brings up the matter of using assistive strategies to deal with precision issues in an adaptive way according to the actual motor skills of each child and not, as most existing touch-based applications for pre-kindergarten children currently implement, in a comprehensive way assuming all children have the same levels of accuracy.

The size of the objects involved in the experiments may have an impact on the effectiveness; hence, future research should be done to evaluate different sizes.

To sum up, the previous results point out that pre-kindergarten children (2 to 3 years old) are in the process of developing their motor skills and have different levels of accuracy. Particularly, older children are able to perform complex gestures with significantly higher levels of accuracy and, therefore, future multi-touch application for this age range should consider assistive strategies that adapt their behavior to the actual levels of motor and cognitive development of each child. Not doing so, would prevent the more skilled children from exercising and further enhancing their precision related skills at an early phase of their development.

**Acknowledgments.** Work supported by the MINECO (grants TIN2010-20488 and TIN2014-60077-R).

## References

1. Abdul, N., Sin, N., Batmaz, F., Stone, R., Chung, P.: Selection of touch gestures for children's applications: repeated experiment to increase reliability. *Int. J. Adv. Comput. Sci. Appl.* **5**(4), 97–102 (2014)
2. Abdul, N., Batmaz, F., Stone, R., Paul, C.: Selection of touch gestures for children's applications. In: Proceedings of SIA 2013, pp. 721–726 (2013)
3. Common Sense Media. Zero to Eight: Childrens Media Use in America (2013)
4. Moser, T., Reikerås, E.: Motor-life-skills of toddlers – a comparative study of Norwegian and british boys and girls applying the early years movement skills checklist. *Eur. Early Child. Educ. Res. J.* **24**, 1–21 (2014)
5. Nacher, V., Jaen, J., Catala, A., Navarro, E., Gonzalez, P.: Improving pre-kindergarten touch performance. In: Proceedings of ITS 2014, 163–166. ACM Press (2014)
6. Nacher, V., Jaen, J., Navarro, E., Catala, A., González, P.: Multi-touch gestures for pre-kindergarten children. *Int. J. Hum.-Comput. Stud.* **73**, 37–51 (2015)
7. Smith, S.P., Burd, E., Rick, J.: Developing, evaluating and deploying multi-touch systems. *Int. J. Hum Comput Stud.* **70**(10), 653–656 (2012)
8. Subban, P.: Differentiated instruction: a research basis. *Int. Educ. J.* **7**(7), 935–947 (2006)
9. Vatavu, R., Cramariuc, G., Schipor, D.M.: Touch interaction for children aged 3 to 6 years : experimental findings and relationship to motor skills. *Int. J. Hum. Comput. Stud.* **74**, 54–76 (2015)

# The 5-Step Plan

## Empowered Children’s Robotic Product Ideas

Lara Lammer<sup>(✉)</sup>, Astrid Weiss, and Markus Vincze

ACIN Institute of Automation and Control, Vienna University of Technology, Vienna, Austria  
`{lammer,weiss,vincze}@acin.tuwien.ac.at`

**Abstract.** When children and adults work together as partners throughout the design process in a collaborative and elaborative manner, children come up with a wide range of creative and innovative ideas. The 5-step plan is a holistic approach that empowers children as robotic product designers. Researchers as well as educators can use the approach to introduce children with different interests to robotics and explore their interests, desires and needs regarding interactive technology like robots. In this paper, we describe the 5-step plan and present our findings on children’s robotic product ideas from three case studies.

**Keywords:** Educational robotics · Robot design · Child-robot interaction

## 1 Introduction

Hendler separates robots for children into the categories toys, pets, interactive displays, educational robotics, and service robotics [1]. Bertel and colleagues [2] further distinguish between educational robotics (or hands-on robotics) and educational service robots: they argue that research within educational robotics has a tradition for participatory design while “the majority of research on educational service robots is still based on highly controlled experiments in lab settings which cause certain limitations to the transferability and applicability of the results to real-world learning environments”. They also imply that Wizard of Oz settings may foster unrealistic expectations and thus cause children to become disappointed when they meet “real” robots. Popular culture also often serves the benchmarks for real-world robot systems and their behaviors [3, 4], which can result in unrealistic expectations, e.g., Veselovská and Mayerová [5] could observe this with children in educational robotics workshops. Eisenberg and colleagues [6] also warn about the effects of ubiquitous educational technologies whereas Resnick and Silverman [7] suggest “choosing black boxes carefully”.

The constraints of the employed technology and the study set-up have an important impact upon children’s perception and expectations towards robots. What kind of future robotic products would children imagine if we circumvented these constraints and studied their ideas in a creative context combining education and design? In this paper, we explore this question and present our approach – the 5-step plan – along with our findings on children’s robotic product ideas from three case studies.

## 2 Related Work on Children, Technology and Design

Motivation and emotions play an important role in learning. Children are driven to grow and assert themselves, as well as to love and be loved. Considering these drivers, adapting learning activities to children's lives and interests [8], and empowering children to learn through play [9] will motivate them. Various design methods consider growth (learning by doing), diversity (not everyone will arrive with the same set of skills) and motivation. One is Learner-centered Design, which should result in the participants' understanding, create and sustain motivation, offer a diversity of learning techniques, and encourage the individual's growth through an adaptable product [10]. Another is Bonded Design, where children participate for a short but intensive time in activities with adult designers. The techniques include brainstorming to generate new ideas, paper-prototyping design ideas both individually and in small groups, and building a consensus on a final low-tech prototype [11]. Additionally, in Cooperative Inquiry children and adults work together as partners throughout the design process in a collaborative and elaborative manner [12]; this leads to empowered children as well as a wide range of creative and innovative ideas [13]. Finally, Garzotto [14] argues that developing technology in an educational context creates a more holistic view (product focus) underlining a number of benefits: collaboration and discussion skills, project/goal oriented attitudes, and capability of reflection and critical thinking (as well as reflecting on technology) for children; and innovative solutions in the way technology can be exploited in an educational setting.

Resnick and Silverman [7] suggest that the best learning experiences occur, when people are actively engaged in designing and creating things. Schrage [15] also underlines the importance of prototyping and argues that prototypes catalyze discussions among designers and potential users. Yet, prototyping is one phase in "creative improvisation" processes embraced by many highly innovative companies to lead their markets [15]. Design Thinking describes such a process where the designer's perception is matched to people's needs with what is technologically feasible and what is a viable business strategy [16]. The design studio is another method where designers work on complex and open-ended problems, use the precedent and think about the whole, use constraints creatively and rapidly iterate their solutions [17].

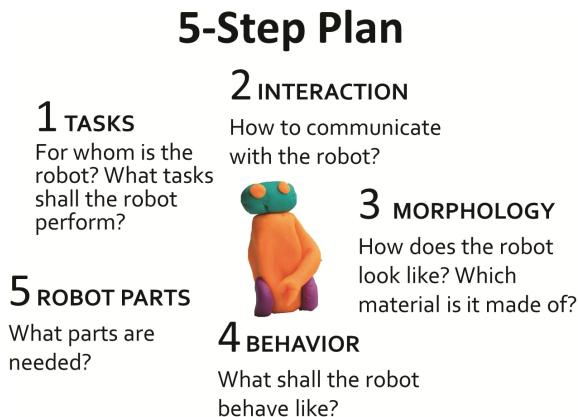
## 3 5-Step Plan

The 5-step plan is an introduction to robotics for children with different backgrounds and varying levels of knowledge. It guides them through the work of real robot experts by introducing them to three important phases of product development: ideation, prototyping and evaluation. Children are encouraged to think as product designers during ideation phase and offered a simple structure to conceptualize a robot from scratch. In this approach, theory and practice are carefully balanced. Children are not constrained by the limits of technology: there are no limits for their robots' capabilities when they start brainstorming. What they cannot build, they describe on paper with words or sketches. In this paper, we focus on the conceptual design (ideation phase). In follow-up sessions – addressing

prototyping and evaluation – maker technology can be used to build working prototypes and evaluate them.

The 5-step plan is based on design methods that empower children [13] to address problems that influence their lives. It uses similar methods like Hamidi and colleagues [18] who introduced children from a rural Mexican community to technology by drawing elements from Learner-centered Design, Bonded Design and Cooperative Inquiry. The plan can be integrated into different teaching or research contexts; it can be adapted to different age groups or even to adults who are not familiar with robotics. We have employed the 5-step plan in three different workshops with children aged 7 to 14.

Figure 1 shows an overview of the 5-step plan as presented to the children.



**Fig. 1.** The 5-step plan

### 3.1 Pre-phase

The researchers introduce themselves as robot experts and explain that in this workshop, the children will learn how to design robots while the researchers will learn from their ideas how to build better robots in the future. As a starting point, children are introduced to technology as “*human-made objects, tools, artifacts that help us*” [19]. We expect that children broaden their view of technology as “*something that we build to make our lives easier*” and start thinking critically about it. The definition of robots draws on the definition of technology: “*True robots act autonomously. They may be able to take input and advice from humans, but are not completely controlled by them*” [20]. We teach children the difference between robots and machines, show them different robotic applications that go beyond public knowledge, and guide them towards the concept of autonomous behavior.

### 3.2 Main Phase

We briefly introduce three incisive stages of a design process (ideation, prototyping, evaluation): *Real robots are highly complex and designed by a team of experts from*

*different disciplines (designers, human-robot-interaction experts, programmers, engineers, etc.). Robot experts consider a few things before they start building prototypes. They ask people, sketch, build models, discuss their ideas, and share them with others. This is what we are going to do today. Each of you will think of a robot idea and then build a model to share it with others.*

Children are guided step by step through five important topics they need to cover if they are to design a robot like a product designer. Pictures help in grasping abstract concepts, but in order to minimize bias on their design ideas the first four steps are explained without any pictures (see also Fig. 1).

*Step 1 – Robot Task (“assignment”).* The children are asked to imagine a robot for themselves that does anything they want. Every idea is valuable in this phase and not discarded as useless or undoable. Children are rather encouraged to think about a helper and adapt their ideas to this concept.

*Step 2 – Robot Interaction.* Known and not yet invented applications are both encouraged equally. Children learn that some of their ideas need scientists who invent new things that are then built into the robots by engineers. *How would you tell your robot what to do? Would you talk to it in a secret language or with signs? Would the robot understand your thoughts? Or would you use an app to control it?*

*Step 3 – Robot Morphology (“looks and materials”).* We have divided the third step, robot morphology, into “looks” and “materials”. First, we introduce four different categories of robot morphology from Fong and colleagues [21]: *Robots can look like machines, like cartoon characters, like animals (zoomorphic) or similar to humans with a head and body (anthropomorphic)*. Second, we talk about different materials robots can be made of, and describe some properties: *They can feel smooth, hard, furry, etc. How would your robot feel like?*

*Step 4 – Robot Behavior.* In the fourth step, the abstract concept of autonomous behavior needs to be explained in a manner that children understand. We use two paths: In order to make the abstract word “behavior” more concrete, we describe roles (or personas) with which children identify. *Would you like your robot to be rather like a butler, a teacher, a protector, a pet or a friend?* We also explain that robots have rules to obey and introduce the Three Laws of Robotics [22]; they offer children a first orientation and are more child-friendly than, e.g., the EPRSC/AHRC principles of robotics.

*Step 5 – Robot Parts.* This last step brings the previous steps together. The researchers show pictures of mechanic and electronic parts: some are used in every robot; others depend on what the robot does, how it looks like or how it should behave. In the beginning of the design process (ideation), the focus is on the holistic view of a product developer who needs to know what parts are needed but is not concerned with the details.

After this introduction children immediately start building a prototype with modeling clay that they can take home to show family and peers. In an expanded 5-step plan concept with follow-up workshops that move from ideation to prototyping, Step 5 is a starting point to go into more detail by using simple technology (e.g. maker electronics) to work out technically feasible solutions.

**Post-Phase.** Once through all five steps, children are encouraged to go back to Step 1 and check if the robot has all parts to accomplish the tasks it was assigned to, then Step 2 to check interaction, then Step 3, etc. Then they start again from Step 1 to check if all fits together. The order of the steps is not as important as the iteration after completing all five steps. When the low-tech prototype is finished, children may present it to the rest of the group. They learn that a robot is developed in iterations and they best start quickly with a simple and cheap prototype to build new ones from the lessons learned.

## 4 Case Studies

The 5-step plan can be integrated into different participatory design or educational robotics contexts. We were involved in three different educational activities:

1. Children University workshop (July 2013, July 2014): one week in summer children aged 7 to 12 attend science and technology lectures at the university and participate in workshops
2. Robot Design workshop (July 2014): two weeks in summer girls aged 10 to 14 participate in science and technology workshops
3. Classroom workshop series (October 2014): five middle school classes (students aged 11–13) each participated in three consecutive science communication project workshops the first of which was “ideation”

Each of these case studies was done with different children in varying contexts and environments in a capital city.

### 4.1 Analysis

We evaluated the 5-step plan from three perspectives: (1) Can children with different interests be empowered to define problems that influence their lives and share their ideas through low-tech prototyping? (2) Does the 5-step plan provide them enough structure for their open creative process? (3) Can researchers derive ideas and needs for future robotic products from these contributions without using the actual technology? In this paper, we report on children’s robot ideas that we derived from the qualitative data from the 5-step plan templates the children had filled out during the workshops. We combined the data of all three case studies and had 114 children in total. The various robot ideas were categorized into meaningful themes, e.g. robots for different types of playing activities were collected individually in the robot for play category. This categorization led to a quantification of the qualitative descriptions, which we analyzed further.

### 4.2 Results: Robots Kids Want

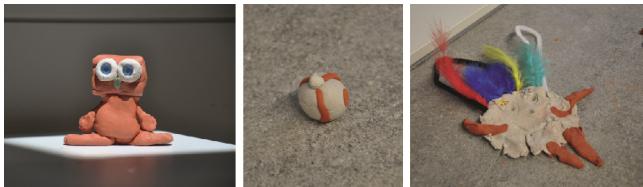
All children had a robot idea to solve a problem from their lives, ranging from cooking robots to protector friends and all of them built a model of their robot from modelling clay. One student even had two ideas and built two models. Many robots (75 %) were related to actual problems out of the children’s lives (including their family), e.g. being

alone at home after school and needing help with cooking or homework, taking care of pets, having entertainment or a playing partner, waking up, or transporting from A to B. The other robots were for people with special needs or special interests (themes with which the children concerned themselves).

In our analysis, we only looked at robots for children that were related to the children's own problems (including their family) and where templates were sufficiently completed, using 83 templates from 38 girls (46 %) and 45 boys (54 %) aged 7 to 14. Table 1 shows which tasks robots should accomplish for children (and their families). Figure 2 shows three examples of children's robot ideas.

**Table 1.** Most mentioned ten tasks and their occurrences.

Task	Occurrence
Play or entertain	23
Do or help with homework	21
Help or serve or both	18
Help in household	18
Cook or serve food	13
Bring or carry or lift objects	12
Talk or make conversation	9
Be a friend	8
Protect	8
Play music or sing	7



**Fig. 2.** Examples (left to right) FrühlingS1000 as good friend who sings and cooks (girl, 13), ETI 5 as reusable exploding play robot (boy, 9), pizza-shaped Cronk to tidy up room (boy, 10)

Interaction with speech was mentioned 56 times, followed by mobile phone or tablet app combined with touchscreen (15 occ.), and then keyboard, mind and remote control with seven occurrences each. The most preferred morphology was anthropomorphic (34 occ.), followed by zoomorphic (21 occ.), machine-like (14 occ.) and cartoon-like (11 occ.). In total, 35 children explicitly stated that their robot should be nice or friendly. This statement was very distinctive. We interpret this in two ways: (1) nice and friendly behavior is an important topic in children's lives; (2) these findings affirm Fong and colleagues [21] identification of mostly "benign" social behavior in social robotics, hence social robots usually designed as assistants, companions, or pets. We also saw this in the personas dedicated to the robots: the 51 children who named personas (not describing adjectives) mentioned friend (12 occ.) and butler (11 occ.) most often,

followed by pet (7 occ.) and butler & friend (4 occ.). Other mentioned personas ( $\leq 3$  occ.) were different combinations of butler, friend, protector and pet (in sum 13 occ.), and teacher (2 occ.) or play partner (2 occ.).

## 5 Conclusion

We have introduced the 5-step plan, an approach that (1) empowers children with different interests to work on technology as product designers; (2) provides them with enough structure for their creative processes; and (3) offers researchers a tool to examine children's robot ideas. The 5-step plan is useful for researchers as well as teachers to introduce all children to robotics, not only children interested in becoming engineers or scientists. We contributed our findings about what types of robots children want to the community to demonstrate that the approach can be used to explore the ideas and needs of a wide range of future robot users.

We have already conducted follow-up workshops in the Classroom workshop series where we used simple electronics, so that interested students could work on engineering tasks while the others could choose other topics of their interest, like marketing, design, or human-robot interaction. We report on this in [23]. We have also started a next round of the Classroom workshop series with four new classes. In these workshops, we have introduced the storyboard technique to analyze how well storyboard and 5-step plan complement each other.

**Acknowledgments.** This project was partly funded by the FWF Science Communication Project "Schräge Roboter" WKP12 and WKP42. We greatly acknowledge the financial support from the Austrian Science Foundation (FWF) under grant agreement T623-N23, V4HRC.

## References

1. Hendl, J.: Robots for the rest of us: designing systems "out of the box". In: Druin, A., Hendl, J. (eds.) *Robots for Kids: Exploring new technologies for learning*. The Morgan Kaufmann Publishers, San Mateo (2002)
2. Bertel, L.B., Rasmussen, D.M., Christiansen, E.: Robots for real: developing a participatory design framework for implementing educational robots in real-world learning environments. In: Kotzé, P., Marsden, G., Lindgaard, G., Wesson, J., Winckler, M. (eds.) *INTERACT 2013, Part II. LNCS*, vol. 8118, pp. 437–444. Springer, Heidelberg (2013)
3. Feil-Seifer, D., Matarić, M.J.: Human–robot interaction (HRI). In: Meyers, R.A. (ed.) *Encyclopedia of Complexity and Systems Science*, pp. 4643–4659. Springer, New York (2009)
4. Kriz, S., Ferro, T., Damera, P., Porter, J.R.: Fictional robots as a data source in HRI research: Exploring the link between science fiction and interactional expectations. In: *Proceedings of RO-MAN*, pp. 458–463 (2010)
5. Veselovská, M., Mayerová, K.: Pilot study: educational robotics at lower secondary school. In: *Constructionism and Creativity Conference*, Vienna (2014)

6. Eisenberg, M., Eisenberg, A., Buechley, L., Elumeze, N.: Invisibility considered harmful: Revisiting traditional principles of ubiquitous computing in the context of education. In: Workshop on Wireless, Mobile and Ubiquitous Technology in Education, WMUTE 2006, pp. 103–110 (2006)
7. Resnick, M., Silverman, B.: Some reflections on designing construction kits for kids. In: Proceedings of Interaction Design and Children Conference, Boulder, CO (2005)
8. Piaget, J., Inhelder, B.: The Psychology of the child (1969)
9. Montessori, M.: The Montessori Method (George, A.E., trans.). Schocken, New York (1964)
10. Soloway, E., Guzdial, M., Hay, K.E.: Learner-centered design: the challenge for HCI in the 21st century. *Interactions* **1**(2), 36–48 (1994)
11. Large, A., Nesset, V., Tabatabaei, N., Beheshti, J.: Bonded Design revisited: Involving children in information visualization design. *Can. J. Inf. Libr. Sci.* **32**(3/4), 107–139 (2008)
12. Druin, A.: The role of children in the design of new technology. *Behav. Inf. Technol.* **21**(1), 1–25 (2002). Taylor & Francis
13. Fails, J.A., Guha, M.L., Druin, A.: Methods and techniques for involving children in the design of new technology for children. *Found. Trends Hum.-Comput. Inter.* **6**(2), 85–166 (2002)
14. Garzotto, F.: Broadening children's involvement as design partners: from technology to experience. In: Proceedings of Interaction Design and Children, pp. 186–193 (2008)
15. Schrage, M.: Serious Play: How the World's Best Companies Simulate to Innovate. Harvard Business School Press, Boston (2000)
16. Brown, T.: Design thinking. Harvard B. Review, June 2009
17. Kuhn, S.: The software design studio: An exploration. *IEEE Softw.* **15**(2), 65–71 (1998)
18. Hamidi, F., Saenz, K., Baljko, M.: Sparkles of brilliance: incorporating cultural and social context in codesign of digital artworks. In: Proceedings of Interaction Design and Children, pp. 77–84 (2014)
19. VDI-Richtlinien. VDI 3780:Technikbewertung. Begriffe und Grundlagen, S.2 (2000)
20. Matarić, M.J.: The Robotics Primer (Intelligent Robotics and Autonomous Agents). The MIT Press, Cambridge (2007)
21. Fong, T., Nourbakhsh, I., Dautenhahn, K.: A survey of socially interactive robots. *Robot. Auton. Syst.* **42**(3), 143–166 (2003)
22. Asimov, I.: Robot. Gnome Press (1950)
23. Lammer, L., Hirschmanner, M., Weiss, A., Vincze, M.: Crazy robots: an introduction to robotics from the product developer's perspective. In: 6th International Conference on Robotics in Education RIE 2015, Yverdon-les-Bains, Switzerland (2015)

# Using IMUs to Identify Supervisors on Touch Devices

Ahmed Kharrufa<sup>1()</sup>, James Nicholson<sup>1</sup>, Paul Dunphy<sup>1</sup>, Steve Hodges<sup>2</sup>, Pam Briggs<sup>3</sup>, and Patrick Olivier<sup>1</sup>

<sup>1</sup> Culture Lab, Newcastle University, Newcastle upon Tyne, UK  
{ahmed.kharrufa, james.nicholson, paul.dunphy,  
patrick.olivier}@ncl.ac.uk

<sup>2</sup> Microsoft Research, Cambridge, UK  
steve.hodges@microsoft.com

<sup>3</sup> PaCT Lab, Northumbria University, Newcastle upon Tyne, UK  
p.briggs@northumbria.ac.uk

**Abstract.** In addition to their popularity as personal devices, tablets, are becoming increasingly prevalent in work and public settings. In many of these application domains a supervisor user – such as the teacher in a classroom – oversees the function of one or more devices. Access to supervisory functions is typically controlled through the use of a passcode, but experience shows that keeping this passcode secret can be problematic. We introduce SwipeID, a method of identifying supervisor users across a set of touch-based devices by correlating data from a wrist-worn inertial measurement unit (IMU) and a corresponding touchscreen interaction. This approach naturally supports access at the time and point of contact and does not require any additional hardware on the client devices. We describe the design of our system and the challenge-response protocols we have considered. We then present an evaluation study to demonstrate feasibility. Finally we highlight the potential for our scheme to extend to different application domains and input devices.

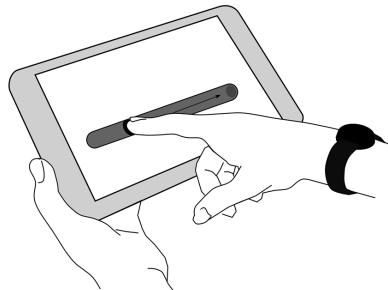
**Keywords:** IMU · Association · Authentication · Touch interaction · UI design

## 1 Introduction

Touch-based computing devices, and in particular a variety of tablet form factors, are becoming prevalent. Initially used as personal devices, they are increasingly being used in work and public settings such as schools, shops, museums and exhibition spaces. The focus in this paper will be on school settings, as a principal example, considering that tablet-based classroom applications are growing rapidly [3] and are seen by many as one of the key classroom innovations of the 21st century. However, the same approach applies to the other settings and for other touch devices such as interactive boards, tabletops, laptops and other devices with interactive screens.

In order to leverage the full benefits that tablets can offer in a classroom setting, a teacher or classroom assistant will often need to override the settings on a student's

device and initiate machine-level or even classroom-level effects such as projecting the work of one student onto the class whiteboard or gaining temporary Internet access [2, 8, 13, 14, 19]. Imagine the following scenario: a student has shown an interesting



**Fig. 1.** SwipeID. Associating touch data with IMU data using a challenge-response protocol.

approach to solving a problem which the teacher wants to share with other students. The teacher initiates supervisor access on that student's tablet, then freezes all the other devices in the classroom and projects to the classroom display for a class discussion. Moving through the class, the teacher and one or more teaching assistants can then authenticate other students' devices to pull content from the board, or to sanction sharing of content between students.

A key requirement to support scenarios like these is the ability to identify and authenticate interactions by supervisory users such as teachers, and to differentiate these from regular users, in this case the students. Here we encounter a known problem – secure authentication is difficult to achieve with any tablet or surface interface, given the ease with which casual observers can engage in “shoulder surfing”. Most commonly, user authentication is based on “something you know” that is meant to be held secret – such as a password or PIN. However, this passcode secret is overly exposed when using touch screen input [15] and can be easily compromised in a public setting such as a classroom [15] where authentication is typically in full view of a number of observers. This inevitably leads to uncontrolled student access to unauthorized applications.

Controlling student access to privileged functions is a very real problem. For example, in 2013 Apple was awarded a \$50 million contract from the Los Angeles School Board of Education, to roll iPads out into public schools across the state – intended to be the first of many such large scale education initiatives. However a year later, the scheme faltered in large part because of identity management and authentication concerns raised when the students found ways to access unauthorized content and applications, leading to significant problems with classroom discipline and ultimately challenging broadband capacity [17, 23, 25].

New authentication and access management solutions in the form of biometrics and near field communication (NFC) may help, but these are only available in some of the latest tablets and there is a pressing need for effective, usable access control for existing

tablets including the estimated 10 million iPads *already* in the classroom in the US alone [17].

In this paper, we offer a novel and elegant authentication solution called SwipeID that simplifies supervisory access to tablets with no NFC or biometric capability. Importantly, it requires no hardware augmentation of the touch device itself. Instead, our solution comprises a wrist mounted inertial measurement unit (IMU) worn only by the supervisor(s), and a challenge-response style interaction protocol involving a very simple set of movements. This setup is depicted in Fig. 1. The aim is to provide a practical alternative when other solutions such as NFC and fingerprint readers are not possible (e.g. for iPads and interactive boards in schools, and for large/fixed touch displays in public settings that do not normally have NFC, fingerprint readers, or cameras). Unlike other options, ours allows for identification at the point of touch – especially useful for large displays like interactive boards. We precede a more detailed description and evaluation of our solution with an overview of related work in the areas of touch and sensor interactions and their deployment in a classroom setting.

Our contribution is as follows: We present a novel system that allows simple, point of contact authentication for any touch screen device. We show how such a system can be used to solve known access management problems when using tablets and other touch screens in the classroom as an example. We demonstrate the efficacy of our system in user studies with four exemplar user interface controls.

## 2 Related Work

In this section we cover previous work which tackles the challenge of identifying users of touch devices. We also briefly review literature relating to the use of movement correlations for building trust, a very relevant topic. Finally we consider previous approaches to the specific challenge of teacher orchestration in the classroom, a key application area which inspired this work.

### 2.1 Identifying Touches

Much work that seeks to identify users on touch devices has focused on the tabletop context. In DiamondTouch [7] sensors are embedded in the chairs of users. In conjunction with a transmitter built into the display it is possible to identify the owner of each touch event. Roth et al. [28] proposed the *IR Ring*; an infra-red (IR) augmented finger-worn ring that transmits IR pulses detectable by cameras embedded into a tabletop. Particular patterns of pulses are unique to particular users, making them suitable for identifying touches across sessions and across different tabletops. A similar approach using a wristband was proposed by Meyer and Schmidt [21]. However, both technologies are only suitable for IR-sensitive optical touch detection systems.

Holz and Baudisch [12] designed a special touch screen to support biometric user authentication based on fingerprint recognition carried out dynamically during each touch interaction. This provides natural per-touch user identification but relies on high resolution, high speed scanning and processing hardware. Harrison et al. [10]

introduced an approach for user identification that relies on sensing electrical properties of the human body when using capacitive touch devices. However, while promising, the experimental results showed that the variability of these electrical properties due to biological and environmental factors can be larger than the variability of such properties between users. Mock et al. [22], on the other hand, explored using raw sensor data from typing on an interactive display for user identification. The system was based on optical touch sensing rather than the more common capacitive touch displays.

HandsDown [30] and MTi [4] are two approaches that rely on handprints for identification against a database of users' hands characteristics rather than fingerprints. Unlike HandsDown, MTi is not limited to camera-based touch sensing. However, for both the touch surface needs to be big enough to accommodate an outstretched hand. Other approaches (e.g. [18, 26]) use an overhead camera for tracking and identifying users. These are most suitable for large fixed displays, in particular tablets.

## 2.2 Inferring Relationships by Sensor Correlation

Researchers have developed a number of techniques to infer the physical relationship between devices being used in conjunction with each other. The use of accelerometers for making associations between multiple devices has been a subject of many investigations [9, 11, 24]. Fujinami and Pirttikangas [9] used the correlation of accelerometer signals in wrist-worn devices and devices embedded in objects to reason about the identity of an object's user. In Smart-Its [11] two objects held together and shaken were associated based on a correlation threshold. Similarly, Patel et al. [24] proposed the use of a shake/pause gesture sequence to pair a mobile device with a public terminal. If the mobile device produces the same shake/pause sequence as that displayed on the terminal, then it is assumed that this is the correct device with which to establish an association. For devices that have already been paired, Chen et al. [6] explored the design space of joint interactions between a smart watch and a smart phone. One of the proposed interactions was to use the accelerometer data from the smart watch to augment the interactions with the phone. Shrirang et al. [31] relied on the correlation between input from a wrist-worn accelerometer and from the keyboard/mouse of a computer terminal to confirm the continuous presence of a user. The user is logged-out in the absence of such correlation.

PhoneTouch [29] used server-side correlation-in-time to allow the use of phones to select targets on an interactive surface using direct touch. One application of associating a detected touch with the phone that caused it, is to use the phones for user identification. However, the need to rely on computer vision to detect phone touches and to distinguish them from finger touches limits its use to vision based touch screens. Moreover, since the system is not specifically designed for user identification, association fails when the system detects more than one touch within the same recognition time frame making it very susceptible to attacks.

### 2.3 Classroom Orchestration

With the emergence of affordable tablet and tabletop computers, there has been an increased interest in deploying large numbers of single- and multi-user devices in classroom environments [2, 8, 13, 14, 16, 19, 20]. The support of teacher orchestration of the classroom has repeatedly been identified as a key challenge [2, 13, 19]. Teachers are often provided with remote monitoring and control tools and the ability to project the content of one of the devices to a large classroom display. This approach may be facilitated by providing the teacher with a dedicated device [2, 16, 19, 20].

However, confining acts of orchestration to a single, static device [2] fails to recognize the real nature of teaching a class, which involves dynamic engagements with the whole group, sub-groups and individuals [27, 32]. Alternative proposals include the provision of orchestration functionalities on a teacher's hand-held device such as a tablet [19, 20]. While this improvement was reported to be useful by one teacher [20], that same teacher described how having to hold and interact with such a device limited their ability to work with the students directly on their tables. The realities of classroom environments result in a wide range of situations where holding a tablet will restrict the quality of teachers' interaction with students. In recognition of such restrictions and the need for teachers to interact directly with the students' devices, the TinkerLamp project [8] used a 'TinkerKey' tangible object with specific visual markers on it. This was automatically identified by the students' tabletops, to allow teachers to issue special commands. However, TinkerKey relies on optical multi-touch sensing and is therefore incompatible with capacitive tablets.

## 3 SwipeID

SwipeID is a system we have developed to support supervisors, such as teachers, as they interact with one or more touch devices. We leverage a wrist-worn IMU connected through Bluetooth to a nearby SwipeID server to identify the teacher's interactions (The server can be any available machine that can connect to both the IMU and the other devices and that can perform simple correlation calculations). A key assumption is that only the supervisors in a particular context are wearing IMUs on their wrist. Each of the touch-based client devices runs the SwipeID client service. The client software allows access to commands and tools that can only be successfully activated by a user wearing a pre-configured IMU.

When a touchscreen user command that requires privileged access is executed, a challenge-response protocol is initiated on-screen, requiring the user to perform a particular sequence of gestures. While interacting with this control, all associated touch data is sent over the network to the server. The server, which is continuously reading IMU sensor data from connected wrist devices, compares this sensor data with the touch data transmitted by the client device and then calculates the correlation between them. If the correlation value is above a predefined threshold, a go-ahead is sent back to the client device. If the correlation is low, a 'reject' is sent back to the client. The client device only communicates with the server when privileged access is requested ensuring that the network is not overloaded during normal usage.

SwipeID has the following properties:

- It allows an arbitrary number of people privileged control of any number of touch-based devices.
- It requires no special hardware on the client devices.
- It removes the need to try and keep a password secret and the need to remember passwords.
- It requires a small and relatively low-cost wrist-worn IMU such as a smart watch for each authorized person, along with a networked machine acting as a server.

### 3.1 Correlating Touch and Sensor Data

During the challenge-response phase, the magnitude of the acceleration of each touch stroke is derived from both IMU and touch data. In the case of the IMU, the data from the on-board accelerometer, gyroscope and magnetometer must be integrated to derive the linear acceleration of the device independent of any confounding movement such as rotation. By using the magnitude of the linear acceleration there is no need to consider the relative orientation of the IMU and touch sensor which could vary with time and depending on how the device is worn. The two data streams are then matched using a correlation-in-time function [9] (Eq. 1) and the resulting probability is thresholded. The touch and IMU data are collected at the same rate, 33 Hz. The data is resampled to account for the intermittent sampling latencies inherent in the devices used. The X and Y touch data is differentiated twice to calculate acceleration. The data is filtered using a 5-sample moving average before and after each derivative calculation, and the magnitude of the acceleration is calculated for both the touch ( $d_1$ ) and linear acceleration ( $d_2$ ) data. The data correlation was calculated as follows:

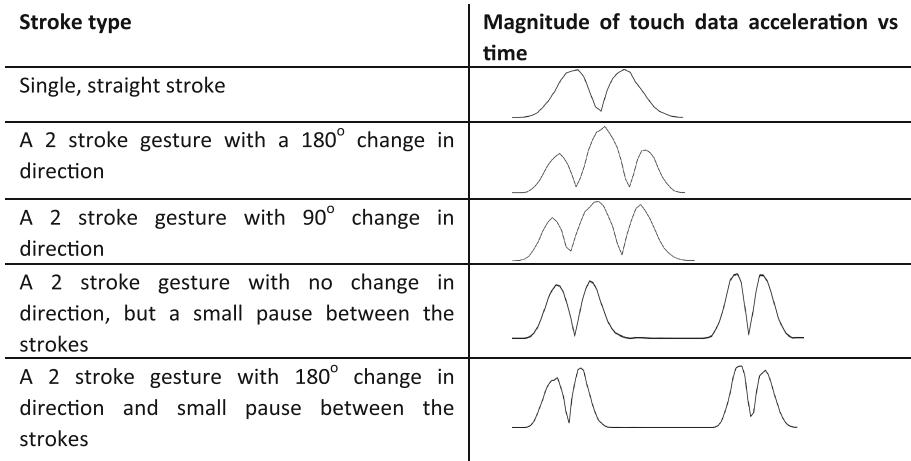
$$r_{12} = \frac{\sum (d_1 - \bar{d}_1) \cdot (d_2 - \bar{d}_2)}{\sqrt{\sum (d_1 - \bar{d}_1)^2 \cdot \sum (d_2 - \bar{d}_2)^2}} \quad (1)$$

### 3.2 A Network Based Challenge-Response Approach

SwipeID relies on the correlation between the acceleration calculated from the touch data and that measured by the IMU. Accordingly, in theory any user could attempt to trigger a privileged command as authentication can only be carried through performing some gestures aiming to achieve the correlation threshold. This means that without imposing constraints on the type of gestures required for authentication, a user not wearing an IMU could attempt to trigger a privileged command concurrently with a supervisor (who is working on another device), and mirror the movement pattern of the supervisor. If the unauthorized user were to do this well enough, there is a possibility that the correlation calculated between their touch data and that of the IMU on the supervisor's wrist exceeds the threshold and might even be higher than that of the

supervisor - due to measuring acceleration at the wrist which does not map perfectly to the touch data at the fingertip.

**Table 1.** The magnitude of acceleration from touch data for basic gestures.



To prevent such attacks, a challenge-response approach is used. The server has a pool of distinct challenges equal to or larger than the number of devices in use at any one time. If two users try to gain privileged access at the same time, they will be assigned two different challenges that require different movement patterns. If unauthorized users try to copy the movement pattern of the person wearing the IMU, they will have to deviate from the movement pattern required by their own challenge and thus fail their challenge locally regardless of how well their movement correlates with that of the IMU. Accordingly, the only option users have to gain access to privileged commands is to both accurately follow their own challenges by wearing the IMU.

To best design a set of challenge patterns that ensure distinct movement patterns which will not inadvertently cause high correlation between different challenges, we looked at the basic acceleration signals generated from touch data for the most primitive strokes (Table 1). From these graphs we can see that by interleaving short strokes and periods of no movement, it is possible generate a number of distinct patterns. If we represent a movement by M and a pause by P, we can design as many different challenges as desired. We want to keep the sequence short to keep it quick to enter, but to decrease the correlation between the different patterns we decided to choose challenges that differ in at least two segments. Table 2 lists 12 different challenges which meet this criteria and also have a minimum of three movements and one pause.

**Table 2.** Twelve different challenges that combine movements (M) and pauses (P). (See Fig. 2 for example shapes of signals associated with these challenges)

M	P	M	P	M	P
M	P	M	P	P	M
M	P	P	M	M	P
M	P	P	M	P	M
P	M	M	P	M	P
P	M	M	P	P	M
P	M	P	M	M	P
P	M	P	M	P	M
M	M	P	P	M	M
P	M	M	M	M	M
M	M	M	M	M	P
M	M	M	M	P	M

**Designing Sensor-Coupled User Controls.** To require a user to follow a specific challenge, a custom user interface control must display it and verify the response. We would like these controls to:

1. be user friendly, i.e. feel simple, be quick to use and not too demanding;
2. support the proposed challenge-response protocol; and
3. result in sufficient movement to generate high correlation.

To aim for user friendliness (goal 1), we avoided controls that increase cognitive load, such as requiring users to trace a dynamic path that appears incrementally or changes in real time. We limited the design space to only include controls that simply require users to move along a path or to move to clearly marked targets. With such controls, it is possible to enforce a pattern that results in different levels of movement and to verify that the user is following the displayed challenge (goal 2).

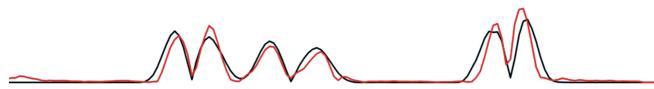
To meet goal 3 we considered two options for imposing distinct changes in the speed of movement:

Present a sequence of targets one at a time. The user is required to move to the next target within a certain time otherwise the challenge fails. Once a target is reached, the subsequent target appears, either immediately or after a certain period, depending on the challenge. This enforces a distinctive movement/no movement pattern. We call this '*discrete*' interaction as there are explicit wait periods. Note that the user is not aware of the full challenge pattern beforehand.

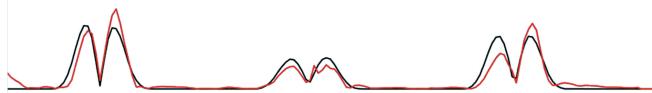
Show the full challenge, represented by a certain path to navigate through, from the outset. The required pattern of movements is achieved by switching between a wide straight paths which can be navigated through quickly, and shorter narrow and/or curvy paths that require more careful (and slow) navigation. This idea is derived from Accot and Zhai's work [1] that looked at the relationship between movement time and width, and to a lesser extent curvature, of a path to steer through. We refer to this as '*continuous*' interaction because no pauses are expected. In this case the users can see the full challenge from the outset.

**Table 3.** Design space for four different IMU-coupled user interface control types. The shapes in the second row are one of 12 different challenge shapes.

	Small footprint	Large footprint
Discrete Challenge		
Continuous Challenge		



A pause-move-move-pause-move-pause challenge (PMMMPMP).



A move-pause-move-pause-move-pause challenge (MPMPMPM).

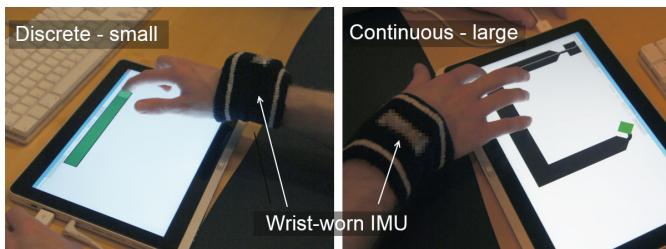
**Fig. 2.** The magnitude of acceleration signal from touch (black) and IMU (red) for two different challenges using the discrete, small footprint control (Color figure online).

It is hypothesized that users associate different levels of ease of use to options (a) and (b) because of the different style of interaction used. In particular, the use of continuous movement and exposure to the full challenge beforehand is assumed to help in perceiving the interaction as one action (or *chunk* [5]) rather than a number of discrete actions. Pauses or very slow movement in a straight wide path area results in failing the challenge. Moreover, as long as the user does not go outside the path they can speed up the response for a continuous challenge by navigating more quickly. This is not the case with the discrete challenge as the user must wait for each target to appear.

We need to ensure sufficient wrist movement to generate meaningful IMU data. We hypothesized that asking users to perform the task as quickly and as accurately as possible would help with this. We also need to avoid finger-only movements. For this reason, we wanted to compare gestures which use predominantly one axis of movement with those which require movement in two dimensions – with both cases requiring movements above a certain minimum physical distance. The full design space is summarized in Table 3.

For discrete, small footprint interactions, a horizontal path is presented with the next target to move to indicated as a circular region. The following target either appears immediately after reaching this target or after a certain pause, depending on the challenge. Similarly, for the large footprint discrete option, a rectangular shape is displayed with the target appearing at one of the four corners.

For continuous interaction, the control is a path that changes between straight, wide segments and narrow curvy segments that the user needs to navigate (steer) through.



**Fig. 3.** User study. Discrete, small footprint control (left). Continuous, large footprint control (right).

The path, which ends with a clearly-marked target, is designed to either stretch mostly along one axis (small footprint) or along both (large footprint).

We envisage privileged commands to be presented as normal buttons that expand once touched. When a small discrete challenge button is touched for example, it expands horizontally to the required width showing the next target. The user must then slide their finger without removing it to the next target and so on. Upon sliding to the last point in the challenge, which is the same as the starting point, the button collapses. The button also collapses immediately if the user fails to follow the challenge.

Figure 2 shows the magnitude of acceleration signal at the server side calculated from touch data and that from the IMU for two different challenges. These signals, which are used for correlation calculations, clearly show the similarity between touch and IMU data and also how different challenges result in different signals.

## 4 User Study

We conducted a study as a proof of concept for SwipeID and to compare the performance and user perception of the four different control types: discrete small footprint, discrete large footprint, continuous small footprint, and continuous large footprint. A repeated measures design was chosen to directly compare participants' performance with each control. We recruited 19 participants for the study (mean age = 26 years; all right handed). Inclusion criteria included good (correctable) vision and some experience with touch-based devices such as smartphones or tablets.

The touch device used was a 10.1" Windows 8 tablet. A commercial IMU, LPMS-B from Life Performance Research was used in the study and was fixed on the

participants' wrists using a sweatband (Fig. 3). This device supports on-board integration of accelerometer, gyro and magnetometer data to calculate linear acceleration. The tablet was placed horizontally on a table and the participants interacted with it from a seated position. All the controls had a physical length of 15.9 cm and the large footprint controls had a height of 12.0 cm.

Initially participants were briefed regarding the purpose of the study and were shown the four control types. They were then given two practice challenges for each control before commencing the study proper. The presentation order for the four controls was counterbalanced and the 12 challenges were randomized for each control (see Table 2 for challenges). Participants were asked to complete each challenge as quickly and accurately as possible. If the participant completed the challenge correctly they were presented with a green feedback screen, whereas a red screen was presented if the challenge was not completed correctly and they were required to retry the challenge until successful. All touch and IMU data from each trial was automatically logged for subsequent analysis.

Upon completing the challenges for each of the four controls, participants were presented with a six-item questionnaire exploring their preferences and perceptions regarding the proposed controls. Four of the questions directly asked the participants to rate their experience using each of the four configurations using a scale ranging from Very Good to Bad. The other questions asked participants to write down their preferred configuration and the configuration they would like to avoid, along with explanations for each selection. Participants were fully debriefed upon completing the questionnaire.

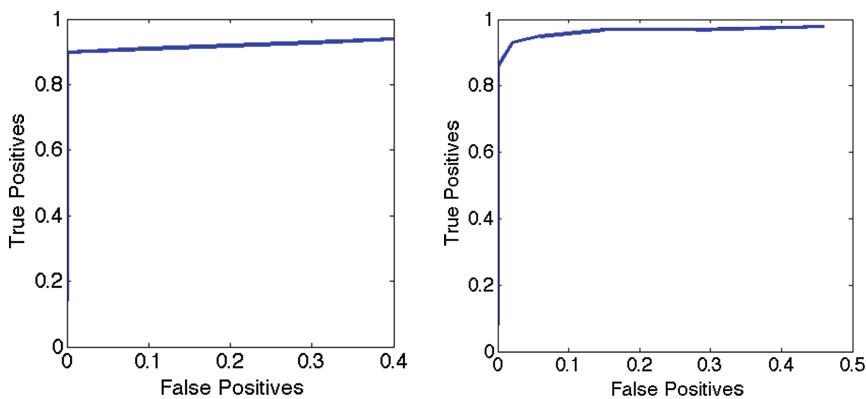
## 4.1 Results

We performed data analysis to determine the efficacy of our challenge-response protocol. To do this we calculated (i) the correlation between corresponding touch and IMU data and (ii) the similarity between touch data for a challenge performed by the person wearing the IMU and touch data from other challenges.

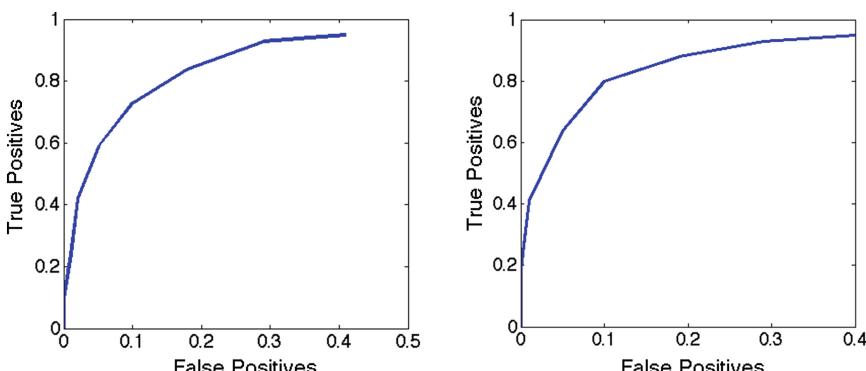
Table 4 shows the average correlation between touch data and IMU data for the different control types as well as the average time to respond to a challenge and the average number of failed attempts to respond to the challenge out of the total of 12 trials. Failure occurs when the user does not follow the challenge accurately, leading to rejecting the response locally without the need to send data to the server. Upon failure, the user had to repeat the challenge. The table shows that the discrete controls resulted in higher average correlation, shorter average time to completion and a lower number of failed attempts on average than the continuous controls. The receiver operating characteristic (ROC) curves for the four controls (Figs. 4 and 5) show that the discrete controls have better performance than the continuous controls. The ROC curves show that for the discrete controls a threshold of 0.6 allows for 86 % (large footprint) to 88 % (small footprint) true positives and 0.2 % (small footprint) to 0.4 % (large footprint) false positives. As for the continuous controls, while the curve still shows that the controls perform well in separating true positives from false positives, a threshold of 0.5 is probably the best choice, which results in 59 % (small footprint) to 64 % (large footprint) true positives but also allows for 5 % false positives.

**Table 4.** Mean (and SD) for correlation, execution time and number of failed challenges across the four different control types proposed.

Control type	Mean correlation	Mean time in msec	Mean no. of failed challenges
Discrete, small footprint	0.77 (0.08)	4800 (435)	1.58 (1.77)
Discrete, large footprint	0.71 (0.07)	5280 (362)	1.79 (2.37)
Continuous, small footprint	0.49 (0.11)	7030 (1800)	6.47 (5.47)
Continuous, large footprint	0.53 (0.07)	7130 (1980)	5.53 (3.01)



**Fig. 4.** Discrete small footprint and discrete large footprint receiver operating characteristic curves.



**Fig. 5.** Continuous small footprint and continuous large footprint receiver operating characteristic curves.

The results show that participants were faster and failed less challenges when using the discrete small footprint controls, while the mean correlation was also the highest. The discrete large footprint control ranked second overall in the three categories, with continuous small footprint third. The continuous large footprint control performed the worst of all the conditions.

The results also show that the large footprint with horizontal, vertical and diagonal movements did not clearly result in better performance than the small footprint with mostly horizontal movements. In the discrete controls case the small footprint performed marginally better than the large footprint control in terms of correlation, time, and failed challenges. In the continuous controls case the large footprint control performed slightly better in terms of correlation and errors, but worse with regards to time. For the discrete control case, one reason why the large footprint control performed slightly worse than its small counterpart may have been the result of participants having to sometimes lift their wrist to see if the next target had been obscured by their hands. This could have led to unwarranted wrist movements and thus IMU acceleration that did not correspond to any touch movement. Such a scenario would result in a reduced correlation between the touch and the IMU data. For the continuous interaction-large footprint control, participants were tracing a path gradually thus no such wrist movements were observed. The higher average correlation in the continuous interaction-large footprint control compared to its small counterpart may be due to the fact that the horizontal only movement could be performed with more finger movement and less wrist movement as compared to a gesture performed in both horizontal and vertical dimensions.

## 4.2 Participant Feedback

All the participants were asked to answer a simple questionnaire to provide subjective ‘experience ratings’ regarding the four control types used in the study. Participants were asked to rate their experience of using each of the four controls on a scale from 1 (bad) to 4 (very good). Additionally, participants were asked to select the control they would prefer to use on a daily basis and also indicate the control they most wanted to avoid. Below we present the findings from the questionnaire.

After aggregating the participants’ experience rating scores for each of the four techniques, we found that the ratings for three of the controls were very close. The discrete small footprint came top with a rating of 3.3, while the discrete large footprint and continuous small footprint controls shared a score of 3.2. The continuous large footprint technique received the lowest rating: 2.8.

Despite the discrete small footprint control receiving the highest experience rating, the most popular control amongst participants (the one most participants selected as their favorite for daily use) was the discrete large footprint technique (see Table 5). This control was perceived to be fast whilst yielding a low number of errors, but participants also noted, as we have observed as well, that some angles could be obscured at times by the placement of the hand (e.g. when target is located in lower-right hand corner) and waiting for the dot to move could become tedious. This may help explain why, despite being the preferred technique by some, it also had a

**Table 5.** Summary of participants' control preference ratings

Control	Preferred choice (%)	Avoid (%)
Discrete small footprint	31.6	10.5
Discrete large footprint	36.8	21.1
Continuous small footprint	31.6	10.5
Continuous large footprint	0	57.9

higher avoidance percentage than both the discrete and continuous small footprint controls by others. The general consensus regarding the discrete small footprint control, the tied-second most popular technique, was that this technique was fast, easy and intuitive.

With regards to the continuous controls, participants agreed that the small footprint technique was fast, but some argued that it was potentially too much work for identification. However, while the small footprint control shared almost the same preference rating as both types of discrete challenge controls, the large footprint version did not obtain any preference votes, and received the majority of avoidance votes. Users mentioned that the technique required too much concentration (i.e., was too much work) and caused too many errors resulting in a 'frustrating' experience. This was a somewhat surprising considering that it had slightly lower failure rate and slightly higher average correlation than the small footprint version, although the average completion time was slightly longer.

The results of our questionnaire show that despite quantitatively being the fastest, leading to fewer errors, and being the highest rated control in terms of experience, the discrete small footprint control was only the second most popular choice amongst participants. However, it is unclear why this was the case as there was no reported negative feedback on this technique. It may be a case of participants' perception of the large footprint being more suitable for the identification task. Clearly, however, participants did not favor the continuous large footprint technique.

## 5 Discussion and Future Work

SwipeID can identify any user wearing the IMU on any touch-based device without the need for special hardware on the touch device, and at the point of contact on the screen. This approach can be a practical solution for devices that do not have a fingerprint or NFC reader and, most immediately, we envision it as a practical solution for supervisor authentication to iPads already in classroom use. Accordingly, we view the decision of using SwipeID as one of practicality rather than based on performance measures alone when compared to NFC, biometric, or password authentication. For example, with devices that do not have NFC or biometric, password is the most likely alternative but, as we have discussed earlier, passwords are not as practical with touch devices in public settings as they are more prone to shoulder surfing especially when it is not possible to use the device in a private environment. Moreover, even with devices that do have NFC or biometric identification, SwipeID could be useful in scenarios where identification is required at a specific point on a large screen where it may be

impractical to move to a specific location for biometric or NFC identification. This includes collaborative work on a large classroom whiteboard where passcode-based authentication procedures are even more visible and could be easily compromised, or when collaborating around digital tabletops for example where issues of reach caused by the large surface render other techniques impractical. With SwipeID authentication is done at the point of contact using simple gestures. The combined use of a challenge-response protocol and our special user control ensure that only the user wearing the IMU is able to successfully respond to the challenge. While this approach does have a certain level of technical complexity, from users' perspective it is a simple technique that is unobtrusive and makes only few assumptions on the user's part.

One of the motivations for this work was to improve the ways in which teachers can interact with students and their touch-based devices in a classroom setting. We note that SwipeID offers a number of opportunities in this space, given that it allows for freedom of movement around the class and allows for device or screen-specific authentication at the point of contact. It also removes the need for special hardware on the students' devices, the use of a secret passcode, or having to issue commands through a separate fixed or handheld computer. As we noted in the introduction, tablets are increasingly used in the classroom, but their uptake is currently being limited by known access management concerns. Better tools for user identification are therefore an important educational issue. However, considering that even with discrete controls, authentication gestures took 4–5 s, this may be perceived to be too long in a classroom context. Exploring the use of SwipeID in a real classroom setting and exploring options for reducing the required gesture time for authentication are two important areas for future investigation.

SwipeID can also be used in other scenarios where supervisor access of touch-based interactions needs to be identified. For example, in a retail context a sales assistant may need to configure the information displayed in-store, while customers are only allowed to browse the information. Alternatively, in a museum a staff member might need to dynamically update information displayed on interactive terminals. Moreover, SwipeID lends itself well to devices with large or fixed interactive displays where it is not possible to interact with the display privately without being prone to shoulder surfing.

In some applications, multiple users share touch-based devices or need to interact collaboratively on a large interactive surface. In these cases it can be beneficial to identify each specific individual in a way that is resistant to shoulder-surfing – again, a known problem with traditional authentication on touch screen devices. If each user is able to use a wrist-worn motion sensor then SwipeID can be used to identify them.

Another possibility for future work is to explore the use of SwipeID with mouse interactions for user authentication (rather than just verifying user presence as in Shrirang et al. [31]). This could allow the use of a single system as a universal authentication solution across a full range of devices within a particular context – educational or otherwise. However we should note that a key challenge with using a mouse is that the user can perform relatively large gestures on the screen without a significant movement of the wrist. The level of correlation between the physical movement of the mouse and the movement of the mouse pointer is also highly dependent on the gain setting of the mouse so this would also need to be factored in.

Our user study aimed to demonstrate the validity of our approach and to evaluate the performance and perception of four proposed controls. It showed that discrete controls performed better than continuous controls in all regards. However, we observed that when users failed a challenge and had to repeat it for the continuous controls case, they performed the gesture faster which resulted in better correlation for that challenge. This appears to support the theory that it is possible for users to improve their performance over time with continuous controls, giving them a longer term advantage over the discrete controls. In future work it may be possible to further optimize the continuous controls to improve performance.

A limitation of our study was that it was conducted in a lab environment rather than in one of the contexts within which we claim its utility. Future work needs to investigate its longitudinal use in an ‘in-the-wild’ environment where supervisors will be interacting with devices of different form factors and from different seating and standing positions. A longitudinal, in the wild evaluation will also help in gaining better understanding of the user’s preferences and may explain some of the discrepancies between preferences and performances recorded in Tables 4 and 5. It can also show whether the measured failure rate can improve with repeated use and whether the level it settles at can cause annoyance to users or not.

SwipeID user identification is reliant on a wrist worn IMU. In other words, the framework identifies the IMU and not the actual user. This means that whoever has the IMU will have privileged access regardless of the actual identity of the user (relying on what the user has rather than what the user knows). This means that the main threat for the system is getting access to the IMU by unauthorized users. This threat is common with other techniques that rely on a hardware key for authentication such as NFCs. Designing the system to overcome this limitation was outside the scope of the work reported here, but one possible solution would be to require a secret passcode to be entered into the wrist-worn IMU prior to use. The user would only be required to enter the passcode when the wrist-band with the IMU is first put on, which will normally just be once and could be done discreetly (thus adding the requirement of something the user knows as well). An alternative to the passcode is to take advantage of the motion sensors and use behavioral biometric to continuously authenticate the person wearing the sensor.

With the increasing popularity of wrist-worn devices incorporating motion sensors, such as smart watches, the need for a dedicated wrist-worn sensor could ultimately be eliminated. While our current work uses linear acceleration data from the IMU which is derived from data collected from an accelerometer, a gyroscope and a magnetometer, future work will investigate the use of accelerometer data only, thus eliminating the need for the gyroscope and the magnetometer, reducing cost and increasing compatibility with wrist-worn consumer devices.

## 6 Conclusion

In this paper we considered the problem of identifying supervisors across multi-touch devices and gave an illustrative classroom scenario. We proposed SwipeID, a system that works across any touch device and which provides an IMU to supervisors and

searches for correlations between the IMU and touch events on known devices to identify the display being used by a supervisor. In our user study we explored IMU-coupled user interface controls to accompany the proposed system and found that requiring the user to perform discrete gestures, on small interface controls enabled us to identify the best correlation between touch and IMU data. However, the time required to complete the authentication process may be perceived to be too long in certain contexts, thus reducing authentication time is identified as an area for further investigation. We propose that SwipeID can also be used across multiple devices, that it has significant classroom potential and that future work can explore its applicability with other input techniques and its relevance to other contexts.

**Acknowledgments.** This work was supported by the RCUK Digital Economy Programme-SIDE: Social Inclusion through the Digital Economy EP/G066019/1

## References

1. Accot, J., Zhai, S.: Beyond Fitts' law: models for trajectory-based HCI tasks. In: Proceedings CHI 1997, pp. 295–302. ACM Press, New York (1997)
2. AlAgha, I., Hatch, A., Ma, L., Burd, L.: Towards a teacher-centric approach for multi-touch surfaces in classrooms. In: Proceedings ITS 2010, pp. 187–196. ACM, New York (2010)
3. British Educational Suppliers Association (BESA): Tablets and apps in schools 2013. ICT Series, May 2013
4. Blažić, B., Vladušić, D., Mladenić, D.: MTi: a method for user identification for multitouch displays. *Int. J. Hum. Comput. Stud.* **71**(6), 691–702 (2013)
5. Buxton, W.A.: Chunking and phrasing and the design of human-computer dialogues. In: Baecker, R.M., Grudin, J., Buxton, W.A., Greenberg, S. (eds.) *Human-Computer Interaction: Toward the Year 2000*, pp. 494–499. Morgan Kaufmann Publishers, San Francisco (1995)
6. Chen, X.A., Grossman, T., Wigdor, D.J., Fitzmaurice, G.: Duet: exploring joint interactions on a smart phone and a smart watch. In: Proceedings CHI 2014, pp. 159–168. ACM Press, New York (2014)
7. Dietz, P., Leigh, D.: DiamondTouch: a multi-user touch technology. Mitsubishi technical report (2003)
8. Do-Lenh, S.: Supporting reflection and classroom orchestration with tangible tabletops. Ph.D thesis, École Polytechnique Fédérale de Lausanne (2012)
9. Fujinami, K., Pirttikangas, S.: A study on a correlation coefficient to associate an object with its user. In: Proceedings 3rd IET International Conference on Intelligent Environments (IE 2007), pp. 288–295 (2007)
10. Harrison, C., Sato, M., Poupyrev, I.: Capacitive fingerprinting: exploring user differentiation by sensing electrical properties of the human body. In: Proceedings UIST 2012, pp. 537–544. ACM Press, New York (2012)
11. 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: Proceedings Ubicomp 2001, pp. 116–122. ACM Press, New York (2001)
12. Holz, C., Baudisch, P.: Fiberio: a touchscreen that senses fingerprints. In: Proceedings UIST 2013, pp. 41–50. ACM Press, New York (2013)

13. Kharrufa, A., Balaam, M., Heslop, P., Leat, D., Dolan, P., Olivier, P.: Tables in the wild: lessons learned from a large-scale multi-tabletop deployment. In: Proceedings CHI 2013, pp. 1021–1030. ACM Press (2013)
14. Kharrufa, A., Martinez-Maldonado, A., Kay, J., Olivier, P.: Extending tabletop application design to the classroom. In: Proceedings ITS 2013, pp. 115–124. ACM, New York (2013)
15. Kim, D., Dunphy, P., Briggs, P., Hook, J., Nicholson, J.W., Nicholson, J., Olivier, P.: Multi-touch authentication on tabletops. In: Proceedings CHI 2010, pp. 1093–1102. ACM Press, New York (2010)
16. Kreitmayer, S., Rogers, Y., Laney, R., Peake, S.: UniPad: orchestrating collaborative activities through shared tablets and an integrated wall display. In: Proceedings UbiComp 2013, pp. 801–810. ACM Press, New York (2013)
17. Leonard, D.: The iPad goes to school (2013). <http://www.businessweek.com/articles/2013-10-24/the-ipad-goes-to-school-the-rise-of-educational-tablets>
18. Martinez, R., Collins, A., Kay, J., Yacef, K.: Who did what? Who said that? Collaid: an environment for capturing traces of collaborative learning at the tabletop. In: Proceedings ITS 2011, pp. 172–181. ACM Press, New York (2011)
19. Martinez-Maldonado, R., Kay, J., Yacef, K., Edbauer, M., Dimitriadis, Y.: MTClassroom and MTDashboard: supporting analysis of teacher attention in an orchestrated multi-tabletop classroom. In: Proceedings CSCL 2013, pp. 320–327 (2013)
20. Mercier, E., McNaughton, J., Higgins, S., Burd, E., Maldonado, R.M., Clayphan, A.: Interactive surfaces and spaces: a learning sciences agenda. In: ICLS 2012 (2012)
21. Meyer, T., Schmidt, D.: IdWristbands: IR-based user identification on multi-touch surfaces. In: Proceedings ITS 2010, pp. 277–278. ACM Press, New York (2010)
22. Mock, P., Edelmann, J., Schilling, A., Rosenstiel, W.: User identification using raw sensor data from typing on interactive displays. In: Proceedings Intelligent User Interfaces (IUI 2014), pp. 67–72. ACM Press, New York (2014)
23. Needle, D.: L.A. school district puts the brakes on massive iPad deployment (2014). <http://tabtimes.com/feature/ittech-os-ipad-ios/2014/08/26/la-school-district-puts-brakes-massive-ipad-deployment>
24. Patel, S.N., Pierce, J.S., Abowd, G.D.: A gesture-based authentication scheme for untrusted public terminals. In: Proceedings UIST 2004, pp. 157–160. ACM Press, New York (2004)
25. Pierra, P.: Tablets-in-school galore causes a bandwidth war in Southern California (2014). <http://tabtimes.com/case-study/education/2014/05/15/tablets-school-galore-causes-bandwidth-war-southern-california>
26. Ramakers, R., Vanacken, D., Luyten, K., Coninx, K., Schöning, J.: Carpus: a non-intrusive user identification technique for interactive surfaces. In: Proceedings UIST 2012, pp. 35–44. ACM Press, New York (2012)
27. Reid, D.J.: Spatial involvement and teacher-pupil interaction patterns in school biology laboratories. *Educ. Stud.* **6**(1), 31–41 (1980)
28. Roth, V., Schmidt, P., Güldenring, B.: The IR ring: authenticating users' touches on a multi-touch display. In: Proceedings UIST 2010, pp. 259–262. ACM Press, New York (2010)
29. Schmidt, D., Chehimi, F., Rukzio, E., Gellersen, H.: PhoneTouch: a technique for direct phone interaction on surfaces. In: Proceedings UIST 2010. pp. 13–16. ACM Press, New York (2010)
30. Schmidt, D., Chong, M.K., Gellersen, H.: HandsDown: hand-contour-based user identification for interactive surfaces. In: Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries, pp. 432–441 (2010)

31. Mare, S., Markham, A.M., Cornelius, C., Peterson, R., Kotz, D.: ZEBRA: zero-effort bilateral recurring authentication. In: Proceedings of the 2014 IEEE Symposium on Security and Privacy (SP 2014), pp. 705–720. IEEE Computer Society, Washington, DC (2014)
32. Smith, H.A.: Nonverbal communication in teaching. *Rev. Educ. Res.* **49**(4), 631–672 (1979)

# Design and Usability Evaluation of Adaptive e-learning Systems Based on Learner Knowledge and Learning Style

Mohammad Alshammari<sup>1()</sup>, Rachid Anane<sup>2</sup>, and Robert J. Hendley<sup>1</sup>

<sup>1</sup> School of Computer Science, University of Birmingham, Birmingham, UK  
`{m.t.m.alshammari,r.j.hendley}@cs.bham.ac.uk`

<sup>2</sup> Faculty of Engineering and Computing, Coventry University, Coventry, UK  
`r.anane@coventry.ac.uk`

**Abstract.** Designing effective adaptive e-learning systems, from a usability perspective, represents a challenge because of the complexity of adaptivity in order to meet the diverse requirements of learners. Furthermore, there is a lack of well-designed experimental evaluation of adaptive e-learning systems in general, and of their usability in particular. The aim of this paper is the presentation of an adaptive e-learning system based on learner knowledge and learning style, and of the results of an initial experimental evaluation of the usability of its two modes of operation. This involves comparing the usability of an adaptive version of the system with the usability of a non-adaptive version, in a learning environment with 75 participants. The experiment produced significant results; they indicate that an adaptive e-learning system based on learner knowledge and learning style has a higher level of perceived usability than a non-adaptive e-learning system. This may also increase the level of satisfaction, engagement and motivation of learners and therefore enhance their learning.

**Keywords:** Usability · Adaptivity · Learning style · e-learning · Experimentation

## 1 Introduction

Designing effective adaptive systems, from a usability perspective, is seen as a challenging task [1]. Adaptive e-learning systems tailor instructional material to the learner's needs by, for instance, providing personalized learning paths, changing the interface layout or hiding some material links [2–4]. Meeting the learner's requirements, providing relevant instructional material and supporting the learner-system interaction goals are increasingly important concerns in e-learning systems [2, 3].

Adaptive systems may, however, violate standard usability principles such as consistency, privacy and learner controllability [1, 5]. Eliminating the negative effects on usability is an essential part of the iterative design process of adaptive systems [6]. It is argued that if an e-learning system is not sufficiently usable, learners become frustrated and focus on the system rather than on the learning content [7]. Furthermore, an e-learning system may be usable in terms of its usage but not in terms of the pedagogical perspective. This issue, therefore, may lead to less effective and less efficient learning with these systems. Usability represents a challenge that should be taken into account

when designing and evaluating adaptive e-learning systems [5, 8]. There is a requirement for a better understanding of where adaptation in e-learning systems is useful, and where it is harmful [7].

Although there are many adaptive e-learning systems that have been designed and implemented, they suffer from a lack of experimental evaluation in general [4]. More particularly, usability evaluation is not usually considered as one of the main criteria in the iterative design process of these systems or in determining their ease of use. According to Zaharias “very little has been done to critically examine the usability of e-learning applications” [9]. It is not always clear how easy and pleasant an adaptive e-learning system is to use.

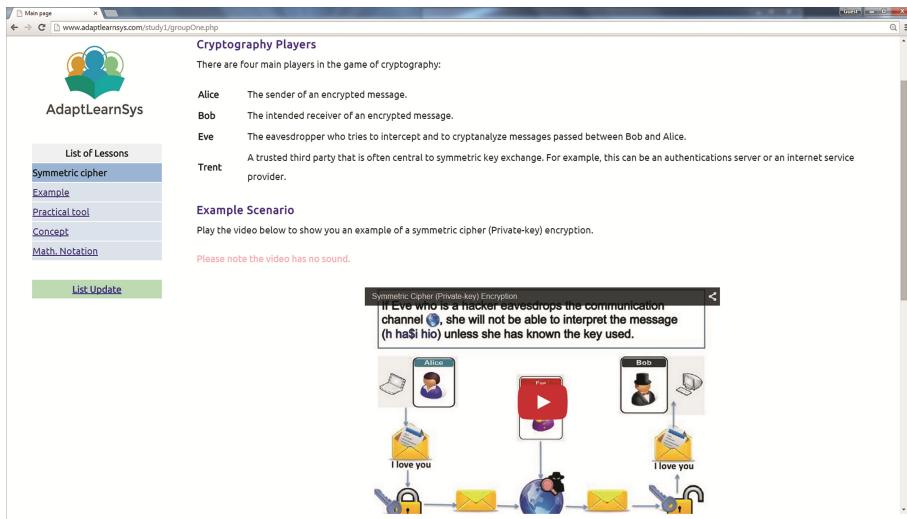
This paper aims to fill a research gap in understanding the usability of adaptive e-learning systems. An adaptive e-learning system, based on knowledge level and learning style, was designed and implemented. It provides personalised learning paths by organising material links according to their relevance to a particular learner; it also provides adaptive guidance and feedback to support learner-system interaction goals. Using a standard usability instrument an experimental evaluation, concerning learners’ perception of usability, was conducted to compare the adaptive e-learning system with a non-adaptive version. The main aim is to determine whether adaptivity influences perceived usability given the fact that both systems have the same interface layout. It is expected that a highly usable e-learning system will increase the satisfaction, engagement and motivation of learners, and therefore, enhance their learning [7, 9]. This study is focused on usability. The evaluation of the learning outcome and learner satisfaction in using the system has been reported elsewhere [10].

The paper is organised as follows. Section 2 presents related work. Section 3 describes the implemented adaptive e-learning system. Section 4 provides the usability evaluation method. Section 5 discusses the usability results, the main usability issues, identifies its limitations and points to future work. Section 6 concludes the paper.

## 2 Related Work

Adaptivity is an approach designed to meet the different needs of different learners when it is incorporated in e-learning systems [2, 3, 11]. Many adaptive e-learning systems have been designed and built [2, 4]. For example, the SQL-Tutor is an intelligent e-learning system that customizes the sequence of SQL lessons based on the knowledge level of learners [12]. An approach that takes into account the learning style of learners to provide instructional recommendations to learners is represented by the eTeacher system [13]. The Protus system combines knowledge level and learning style to personalize learning material for teaching Java programming [14].

Despite some successful systems, the literature also reports some failures [4, 15]. Adaptive e-learning systems provide different designs and presentations to meet the diverse requirements of learners; therefore, usability evaluation plays an important role in the early stages of design, in order to help produce better systems. According to Bangor, Kortum, and Miller “it has become clear that a generalized assessment of the usability and customer satisfaction for different delivery types of interfaces is valuable



**Fig. 1.** A screenshot of the AdaptLearn interface

information to have when trying to determine which technology would be best suited for different deployments” [16]. This highlights the importance of usability evaluations of adaptive e-learning systems in particular, since learners encounter different designs and presentations of content adapted to their characteristics (e.g., learning style). In addition, usability investigations should be taken into account to achieve a harmony between the learner, the learning task, the learning context and the e-learning system [8]. An adaptive e-learning system may be usable in terms of its usage but not in terms of the pedagogical perspective [7].

Usability has a clear connection with learning and also with adaptivity. However, usability evaluation is not usually taken into account; neither when forming the iterative design process of these systems nor in determining how they are easy and pleasant to use [9]. This situation warrants a study of the design and development of an adaptive e-learning systems and of its usability evaluation.

### 3 AdaptLearn: An Adaptive e-learning System

AdaptLearn<sup>1</sup> is an acronym for ‘Adaptive Learning’ which is an adaptive e-learning system. It takes into account learning style and knowledge level as important learner characteristics in order to construct personalized learning paths and adaptive feedback and guidance. A screenshot of AdaptLearn is presented in Fig. 1. The adaptive e-learning framework used as a basis for AdaptLearn, and relevant technical details are presented in our previous work that can be found in [10]. The main components of AdaptLearn are the domain model, the learner model and the adaptation model. They are briefly described below.

<sup>1</sup> AdaptLearn is implemented using NetBeans environment, PHP, JavaScript and MySQL.

List of Lessons Symmetric cipher	List of Lessons Symmetric cipher	List of Lessons Key exchange	List of Lessons Key exchange
<u>Example</u>	<u>Math. Notation</u>	<u>Example</u>	<u>Concept</u>
<u>Practical tool</u>	<u>Example</u>	<u>Concept</u>	<u>Math. Notation</u>
<u>Concept</u>	<u>Practical tool</u>	<u>Math. Notation</u>	
<u>Math. Notation</u>			<u>Example</u>

**Fig. 2.** Examples of personalized learning paths for different learners.

In AdaptLearn, the domain model is structured as a hierachal network (i.e., a tree-like structure) storing knowledge elements related to the application domain. This structure is widely used as a method to representing domain models in related work [17]. The learner model incorporates knowledge level and learning style as learner characteristics in order to provide adaptivity. The knowledge level is an important characteristic that should be taken into account in online learning [2]; it is initialized by using a pre-test, and maintained based on learner-system interaction, mainly test items (i.e., associated with each knowledge element) as a main source of interaction data. The learning style is also assumed to enhance learning when it is integrated in the learner model [18]. It is initialized by using a learning style questionnaire following the Felder-Silverman model which is considered as a valid and reliable learning style identification tool [19].

The adaptation model aims at recommending relevant instructional material to learners to support their interaction goals. It uses the information stored in both the learner model and the domain model in order to provide adaptivity. The adaptation model provides two main adaptive methods including: personalized learning paths and adaptive guidance. The output of the adaptation model is transferred to the interaction model (i.e., the interface).

Personalized learning paths are constructed and continually updated for individual learners based upon their knowledge level and learning style. These paths prioritize links to knowledge elements, hide or remove links to the elements which are not yet ready to be studied (according to the current learning state of the learner), and/or generate links to more relevant knowledge elements, as needed. Examples of personalized learning paths (as provided by AdaptLearn) are presented in Fig. 2. The provision, removal and ordering of elements in the recommended learning paths are expected to meet the needs of learners by taking into account their learning style and their knowledge level. This helps eliminate the effect of cognitive overload, support their interaction goals and optimize the learning process [3].

Adaptive guidance is another form of adaptivity. It is integrated to guide learners as they progress through the learning process to successfully accomplish their learning tasks. It is mainly activated when the learner answers some test items that are related to a specific knowledge element. The system may help the learner review and understand a particularly difficult knowledge element by fetching related supplementary material from the domain model and recommending it to the learner. The system may also highlight some important points for the learner to consider. It may advice the learner to revise a specific knowledge element, and provide general feedback about the learning progress.

## 4 Usability Evaluation

This work contributes to the need to compare different types of interface or system technologies in critical application domains such as e-learning [16]. The learners' perception of usability, using a standard usability instrument, is measured by conducting a controlled experiment in a computer laboratory. A number of experimental sessions were conducted, and each session lasted for up to 2 h. Two experimental conditions were proposed: adaptive condition and non-adaptive condition. In the adaptive condition, a group of subjects interact with an adaptive version of the system. In the non-adaptive condition, another group of subjects interact with the same system but without adaptivity, and a fixed learning path is provided. The main hypothesis that is put forward for this study is that: **an adaptive e-learning system based on learner knowledge and learning style has a higher level of perceived usability than a non-adaptive e-learning system.**

In both conditions, the same interface layout and learning material were used, and the experiment was completed within an equivalent timeframe. Learning material is related to computer security (i.e., the application domain) covering the topics of private-key encryption, public-key encryption and key exchange protocols. Each topic has a number of interrelated knowledge elements. Participants were asked to complete all the lessons provided by the system and to precisely follow its recommendations. The main distinction between the two conditions is the provision of adaptivity.

With regards to the experimental procedure, the participants introduced to the main objectives of the experiment, the system features and informed of the procedure. Second, the participants were asked to access the system via an Internet browser. Then, the participants registered with the system, completed some personal information (e.g., username, age and gender), the Index of Learning Style (ILS)<sup>2</sup> questionnaire based on the Felder-Silverman model [19] and completed a pre-test containing 22 multiple-choice questions to initialize the learner model. Once completed, the system randomly assigned the participant to a specific study condition: the adaptive condition or the non-adaptive condition. The system then directed the learner to the main page to study the learning material. At the end of the learning process, the participant completed the System Usability Scale (SUS) questionnaire [20]. This tool is a quick, reliable and widely used test of system usability [21]. SUS is a 10 item questionnaire with 5-point Likert scale with anchors ranging from "strongly disagree" to "strongly agree".

## 5 Results and Discussion

The experiment was successfully completed by 75 (57 % male, 42 % female) participants. They were undergraduate students in a computer science degree program. The mean age was 22.21 ( $SD = 3.13$ ). The experimental sample involved 39 participants in the adaptive condition and 36 participants in the non-adaptive condition.

The usability of the adaptive e-learning system and the non-adaptive system was measured and compared. The two systems (i.e., the adaptive and non-adaptive systems) in the experiment are based on the same interface layout and learning material with an

---

<sup>2</sup> <https://www.engr.ncsu.edu/learningstyles/ilsweb.html>.

important distinction of the provision of adaptivity. The usability scores for the adaptive system ( $M = 79.46$ ) and the non-adaptive system ( $M = 71$ ) were good and acceptable in general. This implies that both systems are useful and valuable in learning, and the learners found them easy to use. However, the two systems were compared to get deeper insight into their usability, and whether the provision of adaptivity has any impact. Usability may lead to satisfaction, engagement and motivation of learners, and therefore, highly usable systems are expected to improve learning [7, 9].

As there was homogeneity of variance between the study groups as assessed by Levene's Test for Equality of Variances,  $F = .07$ ,  $P = .79$  and data were normally distributed, an independent sample  $t$ -test was conducted in order to compare the two systems using an alpha level ( $\alpha$ ) of 0.01. It was found that there is a statistically significant difference between the general usability score of the adaptive system ( $M = 79.46$ ,  $SD = 13.14$ ) and the non-adaptive system ( $M = 71$ ,  $SD = 13.67$ ),  $t(73) = 2.73$ ,  $P = 0.008$ . The effect size, which provides an objective measure of how important the effect is, was between medium and large ( $d = 0.63$ ). The hypothesis is therefore confirmed, and it can be inferred that the adaptive-learning system based on learner knowledge and learning style has a higher level of perceived usability than the non-adaptive e-learning system.

The experiment, however, was based on two extreme conditions, adaptive and non-adaptive systems; different levels of adaptivity including controllability should be compared with more experimental conditions to get deeper understanding of their effect. However, designing such experiments to evaluate their usability and learning effectiveness represents a significant challenge [15].

In an investigation of the SUS tool items, the most significant difference is that participants would use the adaptive e-learning system more frequently than the non-adaptive system in response to the item "I think I would like to use this website frequently". This is a very important point in usability of systems, which may provide a justification for the adaptivity and recommendation mechanisms provided by the adaptive system. Adaptivity may influence the learners to perceive that the system would help them when needed, provide them with dynamic support according to their characteristics. The recommendations of the system may enhance their intellectual curiosity, satisfaction, and engagement. In contrast, learners may find the non-adaptive system rigid and unresponsive to their needs. They may be less encouraged to use the non-adaptive system as a tool for learning.

Another significant point in relation to the item "I thought this website was easy to use" is that, participants find the adaptive system, in the long run, easier to use than the non-adaptive system. This is a surprising finding because of the inherited complexity of adaptivity. It was expected that the usage of adaptive systems would not be easier than the usage of traditional or non-adaptive systems. Learners may find it more helpful and easier to use a system that provides personalized feedback and recommendations based on their system interaction. For example, the adaptive system suggests to learners what to do if things go wrong, what to do next or what the current situation of the learning process. It may be the case that once learners gain an appreciation of the adaptive system, they may find it easier to use and more useful.

However, the non-adaptive system may seem to be better than the adaptive system in relation to system learnability. This finding was based on the response to the item

"I think I would need the support of a technical person to be able to use this system". It may imply that the non-adaptive system has less problematic issues when it comes to learning how to use the system very quickly. Another point is that, learners who use the adaptive system are more likely to need technical support to be able to initially use the system. Although this finding is not significant, it should be taken into account when designing adaptive e-learning systems.

Although the adaptive and the non-adaptive e-learning systems had the same interface layout, significant results related to the usability of the adaptive system were generated. It is observed that high level of adaptation in e-learning systems enhances usability. The high level of perceived usability may lead learners to be more satisfied, engaged and motivated to use the adaptive e-learning system [7, 9]. It is expected that highly usable adaptive e-learning systems improve learning.

The findings were based on a short-term study with a relatively small number of participants and with few learning resources. With regards to the SUS tool, it was used to test general usability; there is a need for a more specific tool that examines the usability of adaptive e-learning systems in particular. Such a tool would highlight specific usability issues related to adaptivity which will help in designing better systems. In order to improve the adaptive e-learning system, other important factors such as cognitive load, metacognitive skills and affective state may be integrated. Controllability and transparency between the learner and the system can also be investigated.

## 6 Conclusion

This paper has presented an adaptive e-learning system that was designed and built based on knowledge and learning style to support learner-system interaction goals. An initial experimental evaluation of the system usability was conducted with 75 participants in a learning environment; it produced significant results. They indicate that the adaptive-learning system based on learner knowledge and learning style has a higher level of perceived usability than a non-adaptive e-learning system. As usability has an influence on the learner's satisfaction, engagement and motivation when using e-learning systems, learning enhancement is expected when the system is highly usable.

The experiment was based on a short-term study with relatively small number of participants and with few learning resources incorporated in the system. A long-term evaluation with more participants is desirable in future experiments. An experimental evaluation is also being undertaken that builds empirically on the finding of this experiment to investigate learning effectiveness, efficiency and usability when controllability and different levels of adaptivity are provided.

## References

1. Gena, C., Weibelzahl, S.: Usability engineering for the adaptive web. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) Adaptive Web 2007. LNCS, vol. 4321, pp. 720–762. Springer, Heidelberg (2007)

2. Brusilovsky, P., Millán, E.: User models for adaptive hypermedia and adaptive educational systems. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) *Adaptive Web 2007*. LNCS, vol. 4321, pp. 3–53. Springer, Heidelberg (2007)
3. Oppermann, R., Rasher, R.: Adaptability and adaptivity in learning systems. *Knowl. Transf.* **2**, 173–179 (1997)
4. Akbulut, Y., Cardak, C.S.: Adaptive educational hypermedia accommodating learning styles: A content analysis of publications from 2000 to 2011. *Comput. Educ.* **58**, 835–842 (2012)
5. Höök, K.: Steps to take before intelligent user interfaces become real. *Interact. Comput.* **12**, 409–426 (2000)
6. Jameson, A.: Adaptive interfaces and agents. In: *Human-Computer Interaction: Design Issues, Solutions, and Applications*, p. 105 (2009)
7. Ardito, C., Costabile, M.F., De Marsico, M., Lanzilotti, R., Levialdi, S., Roselli, T., Rossano, V.: An approach to usability evaluation of e-learning applications. *Univers. Access Inf. Soc.* **4**, 270–283 (2006)
8. Benyon, D.: Adaptive systems: a solution to usability problems. *User Model. User-adapt. Interact.* **3**, 65–87 (1993)
9. Zaharias, P., Poylymenakou, A.: Developing a usability evaluation method for e-learning applications: Beyond functional usability. *Intl. J. Hum.-Comput. Interact.* **25**, 75–98 (2009)
10. Alshammari, M., Anane, R., Hendley, R.: An e-learning investigation into learning style adaptivity. In: *The 48th Hawaii International Conference on System Sciences (HICSS-48)*, pp. 11–20 (2015)
11. Alshammari, M., Anane, R., Hendley, R.: Adaptivity in e-learning systems. In: *The 8th International Conference on Complex, Intelligent, and Software Intensive Systems (CISIS 2014)*, Birmingham, United Kingdom, pp. 79–86 (2014)
12. Mitrovic, A.: An intelligent SQL tutor on the web. *Int. J. Artif. Intell. Educ.* **13**, 173–197 (2003)
13. Schiaffino, S., Garcia, P., Amandi, A.: eTeacher: Providing personalized assistance to e-learning students. *Comput. Educ.* **51**, 1744–1754 (2008)
14. Klasnja-Milicevic, A., Vesin, B., Ivanovic, M., Budimac, Z.: E-learning personalization based on hybrid recommendation strategy and learning style identification. *Comput. Educ.* **56**, 885–899 (2011)
15. Brown, E.J., Brailsford, T.J., Fisher, T., Moore, A.: Evaluating learning style personalization in adaptive systems: Quantitative methods and approaches. *IEEE Trans. Learn. Technol.* **2**, 10–22 (2009)
16. Bangor, A., Kortum, P.T., Miller, J.T.: An empirical evaluation of the system usability scale. *Int. J. Hum.-Comput. Interact.* **24**, 574–594 (2008)
17. Papanikolaou, K.A., Grigoriadou, M., Kornilakis, H., Magoulas, G.D.: Personalizing the Interaction in a Web-based Educational Hypermedia System: the case of INSPIRE. *User Model. User-adapt. Interact.* **13**, 213–267 (2003)
18. Felder, R.M., Silverman, L.K.: Learning and teaching styles in engineering education. *Eng. Educ.* **78**, 674–681 (1988)
19. Felder, R.M., Spurlin, J.: Applications, reliability and validity of the index of learning styles. *Int. J. Eng. Educ.* **21**, 103–112 (2005)
20. Brooke, J.: SUS-A quick and dirty usability scale. *Usability Eval. Ind.* **189**, 194 (1996)
21. Tullis, T.S., Stetson, J.N.: A comparison of questionnaires for assessing website usability. In: *Usability Professional Association Conference*, pp. 1–12 (2004)

# How Does HCI Research Affect Education Programs? A Study in the Brazilian Context

Isabela Gasparini<sup>1()</sup>, Simone Diniz Junqueira Barbosa<sup>2</sup>,  
Milene Selbach Silveira<sup>3</sup>, Sílvia Amélia Bim<sup>4</sup>, and Clodis Boscaroli<sup>5</sup>

<sup>1</sup> Departamento de Ciência da Computação, UDESC, Florianópolis, SC, Brazil  
[isabela.gasparini@udesc.br](mailto:isabela.gasparini@udesc.br)

<sup>2</sup> Departamento de Informática, PUC-Rio, Rio de Janeiro, RJ, Brazil  
[simone@inf.puc-rio.br](mailto:simone@inf.puc-rio.br)

<sup>3</sup> Faculdade de Informática, PUCRS, Porto Alegre, RS, Brazil  
[milene.silveira@pucrs.br](mailto:milene.silveira@pucrs.br)

<sup>4</sup> Departamento de Informática, UTFPR, Curitiba, PR, Brazil  
[sabim@utfpr.edu.br](mailto:sabim@utfpr.edu.br)

<sup>5</sup> Departamento de Ciência da Computação, UNIOESTE, Cascavel, PR, Brazil  
[clodis.boscaroli@unioeste.br](mailto:clodis.boscaroli@unioeste.br)

**Abstract.** This paper presents a comparative analysis based on two independent studies of Human-Computer Interaction (HCI) education and research in Brazil. The first study was conducted to understand how HCI has been taught in Brazil, via a survey responded by 114 educators and researchers in the country. The second study analyzed the scientific production of HCI in Brazil from a fifteen-year analysis of full papers published on the Brazilian Symposium on Human Factors in Computing Systems (IHC). Our analysis is based on data-driven visual exploration, and it can help to get insights from the data and to identify how HCI research in Brazil relates to our education programs. We believe this kind of analysis can shed some light in the evolution of HCI in other countries.

**Keywords:** HCI education in Brazil · HCI research in Brazil

## 1 Introduction

In 2006, a group of professors and researchers met during the Brazilian Symposium on Human Factors in Computing Systems (IHC) to discuss a syllabus for HCI courses, forming a working group on HCI education [1]. In that year, the working group established a reference syllabus, which has been adopted in undergraduate courses in several Brazilian universities. Since 2010, the Special Interest Group on Human-Computer Interaction (CEIHC) of the Brazilian Computer Society (SBC) has promoted a series of Workshops on HCI Education (WEIHC) alongside the IHC conference. The workshops offer an opportunity for faculty, researchers and students to discuss relevant issues about teaching, educational experiences, teaching materials and HCI curriculum. In 2014, based on previous survey and discussions held in working groups during the WEIHC, and considering the ACM Computer Science Curricula

(2013) and the Curriculum Guidelines for Undergraduate Degree Programs in Information Systems (AIS and ACM 2010), the syllabus described in Silveira and Prates [1] was revised and new syllabi were proposed for Information Systems, Computer Science and Computer Engineering courses [2].

When devising syllabi, a major challenge of education in all areas is that undergraduate students are educated to tackle problems that may not yet exist before they graduate. As such, in order to form good human-computer (HCI) professionals, we believe that professors should expose undergraduate students not only to established HCI topics, but also to some current research topics.

In this paper, we investigate the relationship of topics addressed by HCI research in Brazil with topics of Brazilian HCI education programs. We have based our analysis on two independent studies: the first analyzed the HCI scientific production in Brazil; and the second investigated how HCI has been taught in Brazil.

This paper is structured as follows: Sect. 2 presents works related with HCI education. In Sect. 3, we describe the studies that based our analysis. Sections 4 and 5 discuss the analysis, and Sect. 6, the final considerations.

## 2 Related Works

The first HCI curriculum recommended by ACM SIGCHI was published in 1992 [3]. Almost twenty years later, in 2011, as an attempt to address requests for “guidance and insight into which HCI methods and skills can or should be taught to new graduates and incoming and current staff members” [4], ACM SIGCHI launched a research project “to understand emerging subjects, topics, and challenges in the field of HCI”. The goal of the project is to characterize current HCI education and provide recommendations for academics and practitioners [5].

The project encompasses a review of core and elective HCI courses for both academic students and professionals; a review of textbooks and related teaching materials; and interviews, surveys and focus groups with students, educators, and practitioners worldwide. The survey was first translated to Portuguese and Chinese, and has been taken by over 500 respondents worldwide. It now has also versions in French and Spanish, to broaden the population sample.

As reported in Churchill et al. [5], methods have a prominent role in HCI education, as 8 of the 11 most highly ranked topics (at least 4.46 on a 1-5 scale) were about design or empirical research methodologies, namely: general qualitative research; interaction design; observation; interviews; prototyping (general); analyzing and applying research; general empirical research; and, paper/low-fidelity prototyping.

Churchill et al. [5] hypothesize that methods were highly valued because “In an era of rapidly emerging technologies it is more important to help students develop critical skills, the ability to frame problems and a keen sensibility around what methods to use when than to teach classic principles of desktop design.”

Qualitative and quantitative research methods were viewed as complementary, and not mutually exclusive. There was little consensus, however, as to whether qualitative research methods should be taught within a technical or a social sciences curriculum.

In reviewing the syllabi, Churchill et al. [5] found that all courses covered the following areas: Foundational theory; Current topics; Tools and technologies; Design research methods; Empirical research methods; and Design Studio. In those courses, students were required to complete one of two kinds of projects: Design, build and test an application; and, Conduct and report on empirical research.

From the syllabi review, Churchill et al. [4] uncovered five course patterns:

- Introduction to HCI model = Foundational theory + Design research methods + Design, build and test an application;
- Graduate methods = design research methods and/or empirical research & methods + conduct and report on empirical research;
- Cycle-based survey model = current topics + empirical research methods + conduct and report on empirical research;
- Standard Technology Course model = Tools and technologies + Design, build and test an application;
- Advanced topics in HCI model = current topics + tools and technologies.

In 2006, CEIHC/SBC proposed a reference syllabus (Fig. 1) that closely follows the first pattern (“Introduction to HCI model”), which has been adopted in undergraduate courses in several Brazilian universities [1].

#### **INTRODUCTION TO HUMAN-COMPUTER INTERACTION**

Evolution (history)  
 Areas and disciplines  
 Interface and interaction  
 Quality of use: usability, communicability and accessibility  
 Return on Investment

#### **THEORETICAL FOUNDATIONS**

Cognitive Engineering  
 Semiotic Engineering

#### **HCI EVALUATION**

Overview: what, why and when to evaluate;  
 Observation and monitoring of use  
 Gathering of users opinion  
 Experiments and performance tests (benchmarking)  
 Interpretative evaluation  
 Predictive Evaluation

#### **USER INTERACTION DESIGN**

Interaction styles  
 Style guides and Interaction  
 Guidelines and Interaction Design Standards

#### **HCI DESIGN PROCESS**

Vision of Software Engineering and of HCI  
 Elicitation and Analysis  
 Task modeling  
 Interaction modeling  
 Storyboarding and Prototyping  
 Designing Online Help systems

**Fig. 1.** Topics recommended in the 2006 syllabus.

The ACM SIGCHI research project on HCI Education also uncovered five challenges that HCI educators need to address, related to the following themes [5]:

1. Finding a balance between unity and interdisciplinary - how HCI is characterized as a discipline.
2. Divide between academia and industry - the different skill sets needed to perform in academia and in practice.
3. The role of computer science (CS) in HCI education - in which ways and to what extent CS knowledge contributes to learning HCI.
4. A standardized or flexible curriculum - addressing multiple areas of practice may enable students to go to the industry and work on a variety of projects.
5. Breadth and depth in HCI - a broad curriculum may lead to diverse ideas from multiple perspectives, and also to a more unified theoretical perspective, whereas depth is mainly associated with rigor.

As HCI becomes increasingly multidisciplinary, HCI Education starts exploring fields like anthropology and social sciences in search of methods and theories that can inform HCI work, such as a range of qualitative research methods. Moreover, the scores given by professors and students varied widely, which also points to the need of a flexible curriculum that can be adjusted to the students' profiles (for instance, future researchers versus future practitioners).

One way to improve HCI practices is by improving HCI education. In this direction, researches are being conducted on HCI Education to comprehend their status, contents and teaching strategies in different countries. For instance, Graham [6] conducted a survey on the teaching and assessment methods and recommended literature for HCI courses. Yammiyavar [7] brings the evolution of the HCI educational institutions and research collaborations in usability (areas, topics and events) in India, highlighting the need for more researchers in this field. The same reality takes place in Brazil, as described in the next section.

### **3 Independent Studies of the Brazilian HCI Community**

In this section we describe the studies we conducted about the HCI community in Brazil. Section 3.1 presents the research on HCI Brazilian Symposium papers and Sect. 3.2 describes the research on HCI education in Brazil.

#### **3.1 HCI Brazilian Symposium Papers**

The Brazilian Symposium on Human Factors in Computing Systems (IHC) is the major HCI conference in Brazil and it is supported by the HCI Special Interest Group of the Brazilian Computer Society. Starting in 1998, the first four editions of IHC were held in the format of workshops and, since then, the symposium became the primary forum for the regular meeting of the Brazilian community of researchers, teachers and professionals who are dedicated to HCI in Brazil [8]. The first event was held in 2 days with about 70 participants; the current event lasts 4 to 5 days, with about 200 participants.

In a research investigating the scientific production on HCI in Brazil [9], the collection of full papers published at IHC was analyzed.<sup>1</sup> Data were collected from the proceedings of the conferences, made available by their coordinators. Then, a web-based system for data gathering was created, and for each paper the following information was stored: (a) basic data: year of the conference, paper title, language, theme of the conference, the keywords attributed by the authors of the paper, abstract, fields of the ACM template (keywords, categories and general term); (b) information about the author: name of each paper's authors, e-mail, an acronym of the institutions, department/institute, country of the researchers, state (if the participants are in Brazil); and (c) record of each individual reference of the paper. More detail about this research and some preliminary analysis can be found in Gasparini et al. [9].

### 3.2 Surveys on HCI Education in Brazil

In order to understand how HCI has been taught in Brazil, three surveys have been applied in the last years. In 2009 the focus was the HCI courses being taught countrywide [10]. In 2012, another survey was conducted in response to a SIGCHI demand [11]. The last one targeted a broader audience, not taking into account specificities of the Brazilian context. This way, in 2013, aiming to deepen the analysis and, mainly, taking into consideration the specificities of the Brazilian context, we designed and applied a new survey [12], through which we collected information on the HCI courses at different levels (undergraduate, Master's, Doctorate, and others); where they were taught and whether they were mandatory; the course topics; the number of hours dedicated to each topic; the recommended bibliography; and the instructors' profiles.

The results of these surveys were collected through online questionnaires and participants were invited via mailing lists and social networks. Despite the efforts to achieve a multidisciplinary group of participants most of them had a computing-based background.

The respondents were from all Brazilian geographical regions, giving an interesting overview of HCI Education in Brazil, mainly from the perspective of Computer Science courses. Notwithstanding, we cannot make any claims regarding its statistical validity, since the population of HCI professors, researchers, students and practitioners is unknown. Table 1 shows the number of participants in each survey.

Concerning the degree level of the professors/lecturers, the survey applied in 2009 indicates that 55 % were PhDs and 38 % had a master's course, and 67.4 % of the respondents listed HCI as one of their main research areas [10]. The majority of the courses taught (141 in total) were for undergraduate degrees (57 %), whereas 23 % for graduate degrees, 18 % were offered both to undergraduate and graduate programs and 2 % did not answer. Most courses (57 %) were classified as introductory courses, while 28 % were considered advanced and finally 15 % were HCI modules and were taught mainly in Software Engineering and Computer Graphics courses.

---

<sup>1</sup> As this research was presented in IHC 2013, the proceedings of 2014 were not included in the evaluation process.

**Table 1.** Number of respondents in each survey.

Year of the survey	2009	2012	2013
Number of respondents	89	109	114 <sup>a</sup>

<sup>a</sup>Although the 2013 survey had 114 respondents, only 74 were considered valid, i.e., taught at least one course.

In 2013, only 35 % (26 out of 74 valid respondents) claimed to have a Master's, Doctorate, or *lato sensu* degree in HCI (6 responses<sup>2</sup>) or in Computer Science, with an emphasis on HCI (24 responses). The respondents' teaching experience averaged 4.7 years. It reinforces that HCI is still a new field in Brazil, because most of the surveyed HCI lecturers do not have advanced degrees in HCI, but in other areas, such as other emphases of Computer Science. None of the mentioned HCI undergraduate courses (70 in total) requires prerequisites, and most of them are mandatory within their majors: 63 % in Computer Science, 86 % in Information Systems and 71 % in Computer Engineering. However, their study load is still low, 75 h on average, which means one course during the entire undergraduate program, which has on average a total of 3200 h.

Differently from the 2009 and 2013 surveys, one of the aims of the survey applied in 2012 was to investigate the respondents' opinions about challenges related to HCI education. All respondents (professors and researchers, students, and practitioners) agree that the following topics are challenges in the Brazilian context [11]:

- Adopting a common curriculum;
- Advocating the importance of HCI to computer scientists;
- How to approach HCI as a complex interdisciplinary field;
- Sufficient practice in HCI;
- Sufficient theory in HCI;
- Building on previous education to reach mastery; and
- Fostering collaboration between different programs.

The results about the curricula, syllabi and bibliography (recommended literature) obtained by the 2013 survey are compared with the results of the investigation about the IHC papers in the next section.

### 3.3 Study Procedure

In order to analyze the interplay between HCI research and education, we surveyed the full papers published in our national IHC conference series<sup>3</sup> and examined the

---

<sup>2</sup> Note that a respondent could have multiple degrees in different areas, so the total number of responses exceeds the total number of respondents.

<sup>3</sup> We have included the entire full paper collection in our analysis, authored by either Brazilian or international researchers.

responses of the 2013 survey on HCI education, gathering the keywords/topics and bibliography.<sup>4</sup>

We cross-referenced the keywords and topics in the papers and courses, highlighting similarities and differences of occurrences, such as: keywords/topics highly mentioned on both paper and courses; keywords mentioned in papers, but not in courses; and topics mentioned in courses, but not in papers. When analyzing the bibliography, we also noted most frequently cited references and recommended textbooks. Finally, we present a regional distribution of survey respondents and published papers, relating them with the regions where IHC was held.

## 4 Discussion

In this section, we analyze the study results, discussing the course topics and paper keywords (Sect. 4.1), the bibliography recommended in courses and cited in the papers (Sect. 4.2), and an evolution of HCI research in the country (Sect. 4.3).

### 4.1 Topics and Keywords

Considering the keywords present in three or more IHC papers (Fig. 2), we could observe in the first years the predominance of general terms, such as human-computer interaction, interface, usability, ergonomics design (and their variations: interface design, usability evaluation or human-computer interface, for instance). But we could also observe – since the first editions of IHC – the prevalence of a specific keyword, Semiotic Engineering, which is an HCI theory born in Brazil [13]. The growth and maturity of the Brazilian HCI research and researchers could be observed by the specificity and diversity of topics in recent editions of the conference. In Fig. 2, terms are listed in number of appearance; darker squares indicate high density, and lighter squares indicate low density.

To contrast the paper keywords with the topics of Brazilian HCI undergraduate courses, we depict in Fig. 3 the topics that were either mentioned in both courses and papers, or in more than ten courses. From the three most frequently mentioned keywords in the papers (*Semiotic Engineering*, *Usability Evaluation*, and *Accessibility*), we highlight two that also occur frequently in courses: *Semiotic Engineering* and *Accessibility*, which are detailed in Sect. 5.

Considering now the topics mentioned in courses, we find the predominance of more general terms, such as *Evaluation*, *Usability*, and *Interface Design*, with 28, 18, 18 mentions, respectively.

Some topics present in the paper keywords (Fig. 2) were quite specific and did not appear explicitly in the undergraduate courses (Fig. 3), either because they may be embedded in other topics (e.g. *Personas* and *Cultural Issues in Interaction* or *Participatory Design*), or because they are novel or specific research topics that would not fit in introductory courses (such as *Social Networks*). As expected, some topics were

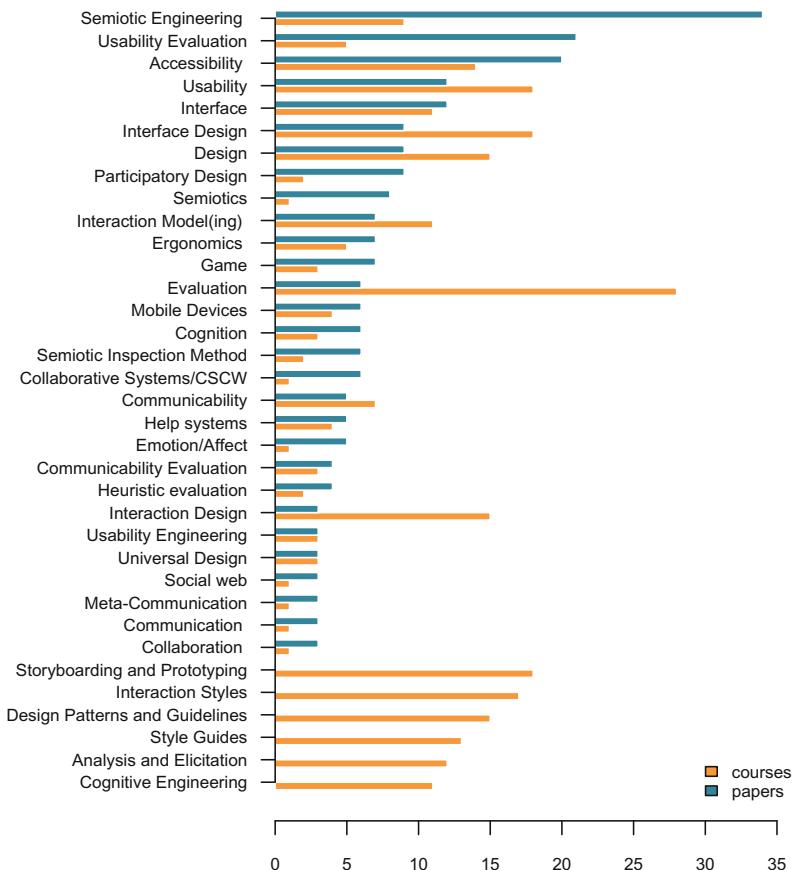
---

<sup>4</sup> For a complementary analysis of the 2013 survey, see [12].

present in undergraduate courses and not in the paper keywords, for instance: *Storyboarding and Prototyping, Interaction Styles, Design Patterns and Guidelines, Style Guides, Analysis and Elicitation, and Cognitive Engineering*.

<b>Keywords</b>	1998	1999	2000	2001	2002	2004	2006	2008	2010	2011	2012	2013
Semiotic Engineering												
Human-Computer Interaction												
Usability Evaluation												
Accessibility												
Interface												
Usability												
Design												
Interface Design												
Participatory Design												
Semiotics												
Ergonomics												
Game												
Interaction Model(ing)												
Social Networks												
Digital TV												
Mobile Devices												
Evaluation												
Cognition												
Distance Education												
User Experience												
Semiotic Inspection Method												
Collaborative Systems/CSCW												
Information Visualization/InfoVis												
Emotion/Affect												
Communicability												
End User Programming												
Adaptive Interfaces												
Human Computer Interface												
Evaluation Methods												
Help systems												
Evaluation of Communicability												
Heuristic evaluation												
Computer-Mediated Communication												
Personas												
Recommender Systems												
Assistive Technology												
Geographic Information System												
Agents												
Cultural issues												

**Fig. 2.** Keywords present in three or more IHC papers



**Fig. 3.** Topics taught in undergraduate courses and present in the papers.

## 4.2 Recommended Bibliography

In the 2013 survey, respondents were asked to present all textbooks, resources and/or references they use in their HCI-related undergraduate courses.

The survey collected 147 different references, mainly books. However, despite the large number of references found, significantly fewer references –only 38– were cited by more than one respondent. Some references were used both in English and Portuguese (because diverse textbooks are translated to Portuguese); other books only in English (they do not have a translation to Portuguese); and still others in Portuguese, written by Brazilian authors. Some references also have different editions.

We compared this set of references to the number of times they were cited in full papers at IHC, in order to explore their similarities. We continue our analysis including only the references that appeared at least three times either in the survey or in the papers (Table 2).

**Table 2.** References cited: Education survey (S) and IHC papers (P).

#	S	P	Lang	Description
1	44	25	EN/PT	PREECE, J.; ROGERS, Y.; SHARP, H. Interaction design: beyond human-computer interaction. John Wiley & Sons, 2013 (+ previous editions)
2	27	6	PT	BARBOSA, S.D.J.; SILVA, B.S. Interação Humano-Computador. Campus/Elsevier, 2010.
3	15	11	PT	ROCHA, H.V.; BARANAUSKAS, M.C.C. Design e Avaliação de Interfaces Humano-Computador. Nied/Unicamp, 2003.
4	14	20	EN	SHNEIDERMAN, B. Designing the user interface. Prentice Hall, 2009. (+ previous editions)
5	9	3	EN/PT	NIELSEN, J.; LORANGER, H. Usabilidade na Web - Projetando Websites com Qualidade. Campus, 2007. (Obs.: translation of "Prioritizing Web Usability")
6	7	5	PT	CYBIS, W.; BETIOL, A.H.; FAUST, R. Ergonomia e Usabilidade: conhecimentos, métodos e aplicações. Novatec, 2010. (+ previous editions)
7	7	26	EN/PT	NIELSEN, J. Usability Engineering. Chestnut Hill, MA: Academic Press, 1993.
8	5	2	PT	DIAS, C. Usabilidade na Web - Criando Portais Mais Acessíveis. AltaBooks, 2007. (+ previous editions)
9	5	10	PT	PRATES, R. O.; BARBOSA, S.D.J. Introdução à Teoria e Prática da IHC fundamentada na Engenharia Semiótica. JAI/CSBC. SBC, 2007.
10	4	0	EN/PT	BENYON, D. Intereração Humano-Computador. Pearson, 2011. (Obs.:2nd edition) (Obs.: translation of "Designing Interactive Systems")
11	4	8	EN	DIX, A.; FINLAY, J.; ABOWD, G.; BEALE, R. Human-Computer Interaction. Prentice Hall, 2004. (Obs.: and/or previous editions)
12	4	8	EN/PT	NIELSEN, J. Projetando Websites. Campus, 2000. (Obs.: translation of "Designing Web Usability")
13	4	3	PT	PRATES, R.O.; BARBOSA, S.D.J. Avaliação de Interfaces de Usuário - Conceitos e Métodos. JAI/CSBC. SBC, 2003.
14	4	7	EN	RUBIN, J.; CHISNELL, D. Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests. Wiley, 2008. (+ previous editions)
15	4	5	EN/PT	NORMAN, D.A. The Psychology of Everyday Things. Basic Books, 1988//The Design of Everyday Things, Basic Book, 1998.
16	3	25	EN	de SOUZA, C.S. The semiotic engineering of human computer interaction. The MIT Press, 2005.
17	3	0	EN	NIELSEN, J.; Tahir, M. Homepage: Usabilidade 50 sites desconstruídos. Campus, 2002. (Obs.: translation of "Homepage Usability: 50 Websites Deconstructed")

(Continued)

**Table 2.** (Continued)

#	S	P	Lang	Description
18	3	0	EN	WILLIAMS, R. Design para quem não é designer: noções básicas de planejamento Visual. Callis, 2005 (Obs.: and/or previous editions)
19	2	24	EN	NIELSEN, J.; MACK, R.L. Usability Inspection Methods. John Wiley & Sons, 1994.
20	2	7	EN	COOPER, A. About face: the essentials of user interface design. New York, NY: Wiley, 2007. (Obs.: and/or previous editions)
21	2	5	EN	The Encyclopaedia of Human-Computer Interaction. <a href="http://www.interaction-design.org/encyclopedia/">http://www.interaction-design.org/encyclopedia/</a>
22	2	5	PT	de SOUZA, C.S.; LEITE, J.C.; PRATES, R.O.; BARBOSA, S. D.J. Projeto de Interfaces de Usuário: Perspectivas Cognitiva e Semiótica. Jornada de Atualização em Informática/CSBC. SBC. 1999.
23	2	3	EN	LAZAR, J.; FENG, J.H.; HOCHHEISER, H. Research Methods in Human-Computer Interaction. John Wiley & Sons, Ltd., 2010.
24	2	3	EN/PT	PRESSMAN, R.S. Engenharia de Software. McGraw Hill, 2002.

Analyzing the data, we can observe that, among the top 10 books cited by educators, only one does not have a version in Portuguese. This reflects a limitation of the Brazilian community regarding teaching material in other languages.

Besides the traditional HCI textbooks, we also see a large number of citations to the seminal Semiotic Engineering book (#16) in research papers. Although it is not frequently cited in undergraduate courses, semiotic engineering is covered at a more basic level by two references that appear in the top 10, namely #2 and #9.

Likewise, Nielsen's and Mack's book on *Usability Inspection Methods* (#19) had only 2 citations at the survey, despite being indicated as a complementary reference in the curricula suggested in 2006 [1]. We believe this happens because the book is written in English, which limits its use by some Brazilian students.

We may also note that the *Accessibility* topic, also suggested by the working group about HCI curricula, is represented in the list mainly by #8, besides the general introductory textbooks.

Considering references in Portuguese, we highlight that references #1 (Preece et al.) and #15 (Norman) have citations both in Portuguese and in English (Table 3).

The need for HCI textbooks written in Portuguese is a demand of the HCI Brazilian community. In the annual event to discuss HCI Education, WEIHC, this topic is one of the most frequently cited, considering the difficulties that our students have in reading material in English [12]. In addition, the government agencies that evaluate our undergraduate courses demand a certain number of titles to our courses (the best concept is associated to a certain quantity of volumes of 3 titles as basic references and 5 titles as complementary references). The need for distinct titles and our students'

**Table 3.** References citation in the undergraduate education survey and in IHC papers, per language (Portuguese and English).

#	Reference	Citation in Survey	Citation in Paper
1	PREECE, J.; ROGERS, Y.; SHARP, H. Interaction design: beyond human-computer interaction. John Wiley & Sons, 2013 (+ previous editions)	34 in Portuguese 10 in English	8 in Portuguese 17 in English
15	NORMAN, D.A. The Psychology of Everyday Things. Basic Books, 1988/The Design of Everyday Things, Basic Book, 1998.	3 in Portuguese 1 in English	0 in Portuguese 5 in English

difficulties with foreign languages increase the search for (and publication of) textbooks written in Portuguese.

In Table 4 we cross-reference the references listed in Table 2 with the symposium year in which each reference was mentioned. The low numbers reflect the fact that textbooks are only rarely cited in research papers.

As we can see, some books that are well known internationally (#1, #4, #7 and #19) appear much more frequently<sup>5</sup> in the conference papers than other books. This is an expected result, because these references are widely recommended and provide a basis to teach and learn diverse HCI topics. Another textbook, *Interação Humano-Computador* (#2), written in Portuguese, has constantly appeared since its first edition in 2010. Other references are more infrequent, and appear in papers whose topics resemble the topics covered by the book.

### 4.3 HCI Around the Country

The IHC conference series is itinerant. Considering the dimensions of Brazil, CEIHC/SBC has held it in distinct regions of the country. The main purpose is to promote HCI research and practice, and to allow community members – professors, researchers, students and practitioners – or anyone interested that lives in the region to participate more thoroughly, which is often very difficult (for example, when the event is held in the South of the country it becomes very costly for those who live in the North or in the Northeast to attend, and vice versa). Specifically related to HCI education, the mobility of the event raises opportunities for HCI faculty (novices or not) to be in contact with the most recent research developed in the country, promoting more advanced discussions in the classroom. WEIHC also provides an opportunity for the local faculty to discuss, to collaborate and to share with the community their classroom challenges. Figure 4 shows the Brazilian states where IHC took place from 1998 to

---

<sup>5</sup> Although some textbooks have a seemingly large number of occurrences in research papers, it is important to notice that we did not count references in all 261 papers published at IHC, but only those which were also mentioned in the educational survey.

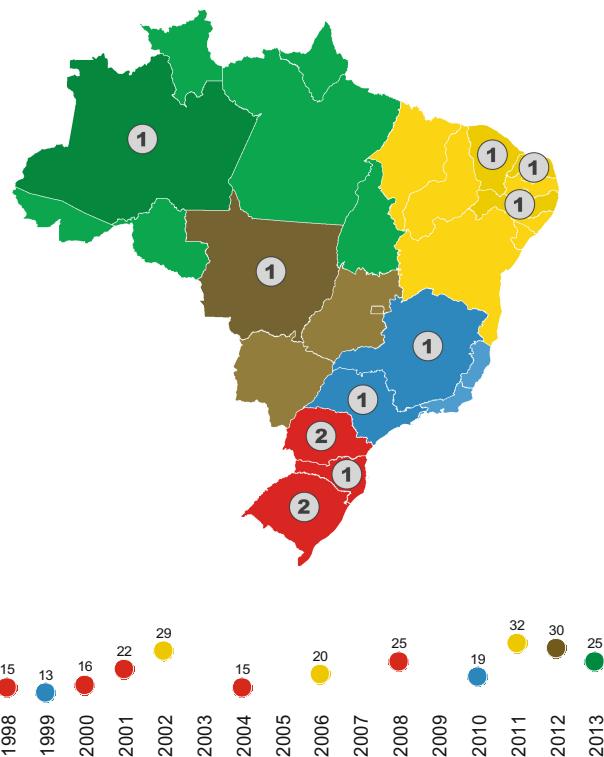
**Table 4.** Number of citations to each recommended book in IHC papers, per year.

#	Reference	1998	1999	2000	2001	2002	2004	2006	2008	2010	2011	2012	2013
1	PREECE, J.; ROGERS, Y.; SHARP, H.				2	1	2	5	3	4	6	2	
2	BARBOSA, S.D.J.; SILVA, B.S.									3	2	1	
3	ROCHA, H.V.; BARANAUSKAS, M.C.C.			1	1		1	3	2	2	1		
4	SHNEIDERMAN, B.	3	3	3	1	2		2	2		4		
5	NIELSEN, J.; LORANGER, H.									2		1	
6	CYBIS, W.; BETIOL, A.H.; FAUST, R.							1	2	2			
7	NIELSEN, J.	2	3	1	4	2	2	3	2	2	4		1
8	DIAS, C.									1	1		
9	PRATES, R. O.; BARBOSA, S.D.J.							2	2	2	3	1	
10	BENYON, D.												
11	DIX, A.		1	3				2			2		
12	NIELSEN, J.			1	1	1			2	1		1	1
13	PRATES, R.O.; BARBOSA, S.D.J.										1	2	
14	RUBIN, J.; CHISNELL, D.				1	1	2			2	1		
15	NORMAN, D.A.			1	2	1			1				
16	de SOUZA, C.S.									2	5	9	9
17	NIELSEN, J.; TAHIR, M.												
18	WILLIAMS, R.												
19	NIELSEN, J.; MACK, R.L.	2		3	4	3	3	2	1		3	2	1
20	COOPER, A.		1						2	1	2	1	
21	The Encyclopedia of Human-Computer Interaction.										2	3	
22	de SOUZA, C.S.; LEITE, J.C.; PRATES, R.O.; BARBOSA, S.D.J.					1	2	1			1		
23	LAZAR, J.; FENG, J.H.; HOCHHEISER, H.										2	1	
24	PRESSMAN, R.S.					2					1		

2013 (the numbers in each state indicate how many times the conference was held there, and the colors represent the five different geographic regions), as well as a timeline with the number of papers published at each event.

As depicted in Fig. 5, we may note that most of the HCI researchers in Brazil who publish at IHC are located in the Southeast region. Figure 6 depicts the evolution of the number of papers per region, over time.

Contrasting Fig. 6 and the timeline depicted in Fig. 4, we see that the number of papers in the regions where IHC is held often increases (see, for instance, increases in the number of papers from the Northeast in 2002, 2006, and 2011; from the South in



**Fig. 4.** States where IHC was held, how many times in each state, and a timeline with the number of full papers published at each event (until 2013).

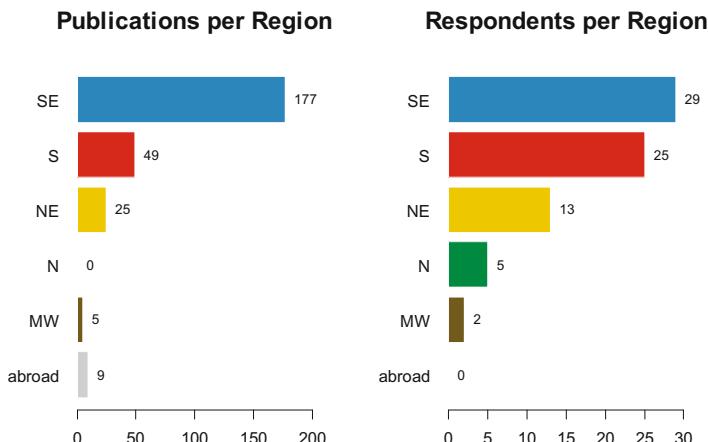
2000 and 2004; and from the Southeast in 2010), which motivates us to keep moving the event around the country.

## 5 Highlighted Cases: Semiotic Engineering and Accessibility

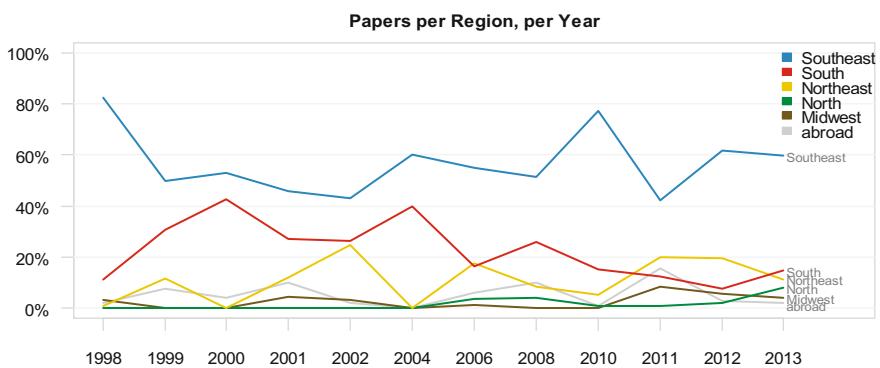
Although we do not have temporal data on the evolution of topics taught in undergraduate classes, we can single out two interesting cases: semiotic engineering and accessibility.

Semiotic engineering [13] is a theory of HCI that was developed in the 90 s at PUC-Rio (in Rio the Janeiro, in the Southeast region of the country) and has been consistently developed over the years and expanded to other geographic states, e.g. one state of the same region (Minas Gerais), states in other regions (South and Northeast), and even from abroad. Figure 7 depicts the states where the authors of the 34 semiotic engineering papers are affiliated.<sup>6</sup>

<sup>6</sup> If a paper is coauthored by three people from different states, it counts as one-third for each state.



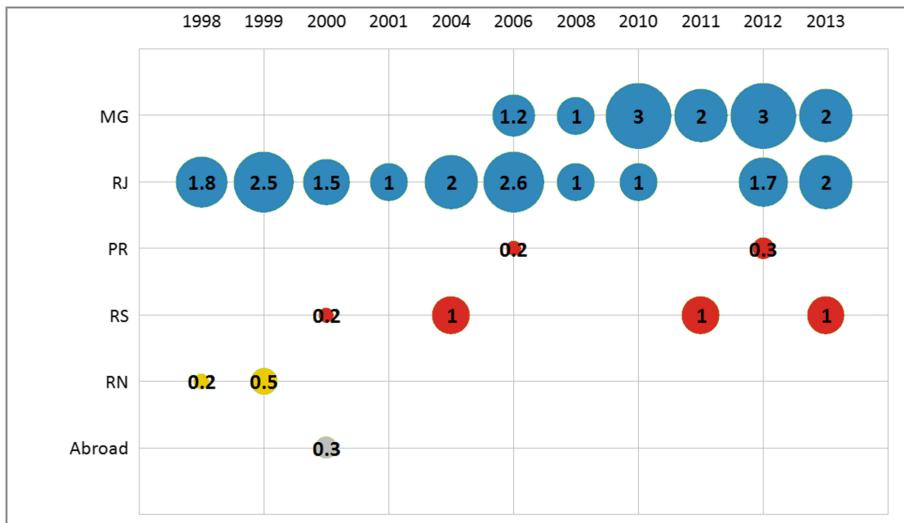
**Fig. 5.** Distribution of publications of the most productive authors (of at least 4 publications) by region, until 2013 (left) and of the respondents of the 2013 survey (right).



**Fig. 6.** Number of papers per region, per year.

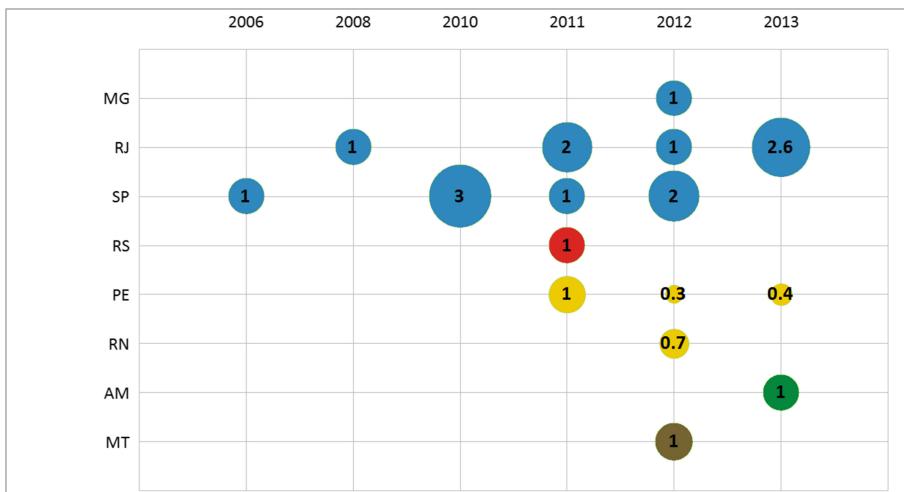
From 2000 on, made its way into both graduate and undergraduate courses, starting at the university where it originated. In 2007, the reference curriculum included semiotic engineering as a topic in undergraduate courses and, in our survey in 2013, it was a topic in 21 courses (including more specific topics, such as communicability), below only general HCI topics, such as interface, evaluation, human-computer interaction, accessibility, usability, and design. The influence of semiotic engineering may also be due to the publication of a variety of books and teaching material throughout the years (tutorials in 1999, 2003 and 2007; research books in 2005 and 2009; and textbook in 2010).

Accessibility is also a notable example of the effects of HCI research on HCI teaching, considering not only the research associated to people with disabilities and the diversity of the technology, but also the growing number of functionally illiterate people in Brazil. The topic is not only a concern of the Brazilian HCI community, but



**Fig. 7.** Semiotic engineering papers over the years, per state and region.

more generally of the Brazilian Computer Society (SBC). In 2006, SBC defined five Grand Challenges in Computer Science Research for the following decade, one of which was: “Participative and universal access to knowledge for the Brazilian citizen” [15]. As several research funding opportunities arose to face the Grand Challenges, more initiatives to investigate issues, methods, and technologies related to accessibility were also formed [16]. Figure 8 presents the 20 IHC papers on accessibility over time, across geographic states of its authors.



**Fig. 8.** Accessibility papers over the years, per state and region.

In the figure, we see that 2006 was also the year when the first papers on accessibility were published at IHC, and in every IHC since then we see an increasing number of papers on accessibility, with a peak in 2012 and an increasing number of states, from different regions of the country.

## 6 Conclusion

From an analysis of the collected data in context, i.e., taking into account the situation of Brazilian research over time, we concluded that the cross-fertilization between research and education has involved at least three factors: (i) the creation of a reference syllabus by an HCI committee sponsored by the country's CS society (in 2006) [1]; (ii) the publication of educational material in Portuguese (books, technical reports, and class notes and slides), including freely available material [12]; and (iii) research funding opportunities promoted by funding agencies [16].

The work presented here is only one of many possible steps in this area. It is important to mention that the majority of the participants of the survey conducted in 2013 in Brazil had a Computer Science background. Consequently, since HCI is a multidisciplinary field, investigations in the context of other disciplines, such as Design and Psychology, could broaden the analysis of the relation between HCI research and teaching in other contexts. For instance, the profile of the participants could explain the low frequency of ergonomics as a topic taught in undergraduate courses. Additionally, most of the participants of IHC are also from Computer Science<sup>7</sup> which could explain the disappearance of the ergonomics as a topic in the papers published in the event.

The relation with international publication topics is also another perspective to investigate, since many Brazilian researchers choose to publish their works in international events in addition to or instead of publishing at IHC [16]. Investigating the international context is also important because nowadays the Brazilian government offers many opportunities for students to stay at least one year studying abroad during their undergraduate courses. It would be important to verify whether Brazilian HCI courses are preparing students for such international experiences.

The investigation and related discussions presented here can help HCI researchers and educators to better understand – and improve – their role in this field. Since its creation in 1998 [17], the Brazilian community has succeeded in promoting the growth of HCI research and practice in the country. We believe that similar analyses can shed some light in the evolution of HCI in other countries.

As future work, we want to conduct new editions of the 2013 survey periodically, so that we can analyze the trends in HCI education in Brazil. We want to explore the networks of collaboration and the researchers' migration within the country. For instance, we want to gather data to evaluate the hypothesis that the number of publications in a region is strongly influenced by the migration of former graduate students from the main research poles (Rio, São Paulo, and Rio Grande do Sul) to emerging

---

<sup>7</sup> For instance, out of the 25 full papers published at IHC 2013, only one was written exclusively by non-CS authors.

universities in various states. We also plan to conduct a study of HCI practice in industry, with the goal to understand the mutual influences between research, practice, and education.

## References

1. Silveira, M.S., Prates, R.O.: Uma proposta da comunidade para o ensino de IHC no brasil. In: XIV Workshop sobre Educação em Computação, vol. 1, pp. 76–84. Sociedade Brasileira de Computação (2007)
2. Boscaroli, C., Silveira, M.S., Prates, R.O., Bim, S.A., Barbosa, S.D.J.: Currículos de IHC no brasil: panorama atual e perspectivas. In: XXII Workshop sobre Educação em Computação no XXXIV CSBC, pp. 1294–1303. Sociedade Brasileira de Computação (2014)
3. Hewett, T., Baecker, R., Card, S., Carey, T., Gasen, J., Mantel, M., Perlman, G., Strong, G., Verplank, W.: ACM SIGCHI Curricula for Human-Computer Interaction (1992). <http://old.sigchi.org/cdg/>
4. Churchill, E.F., Bowser, A., Preece, J.: Teaching and learning human-computer interaction: past, present, and future. *ACM Interact.* **2013**, 44–53 (2013)
5. Churchill, E., Preece, J., Bowser, A.: Information on the 2011–2013 research on HCI Education (2013). <http://www.sigchi.org/resources/education/2011-education-project-1/2011-education-project>
6. Graham, D.A.: Survey of teaching and assessment methods employed in UK higher education programmes for HCI courses. In: 7th HCI Educators Workshop, pp. 66–69. University of Central Lancashire, Preston, UK (2004)
7. Yammiyavar, P.: Status of HCI and usability research in Indian educational institutions. In: Katte, D., Orngreen, R., Yammiyavar, P., Clemmensen, T. (eds.) *Human Work Interaction Design: Usability in Social, Cultural and Organizational Context*. vol. 316, pp. 21–27. Springer, Heidelberg (2012)
8. de Souza, C.S.: Da importância dos Simpósios Brasileiros de Fatores Humanos em Sistemas Computacionais, (2013). [http://comissoes.sbc.org.br/ce-ihc/documentos/da-importancia-dos-IHcs\\_2006.html](http://comissoes.sbc.org.br/ce-ihc/documentos/da-importancia-dos-IHcs_2006.html)
9. Gasparini, I., Kimura, M.K., Pimenta, M.S.: Visualizando 15 anos de IHC. In: Proceedings of 12th Brazilian Symposium on Human Factors in Computing Systems (IHC 2013), pp. 238–247. Sociedade Brasileira de Computação (2013)
10. Prates, R.O., Filgueiras, L.: Usability in Brazil. In: Ian Douglas; Liu Zhengjie. (Org.). *Global Usability*, 1ed. vol. 1, pp. 91–110. Springer-Verlag, London (2011)
11. Boscaroli, C., Bim, S.A., Silveira, M.S., Prates, R.O., Barbosa, S.D.J.: HCI education in brazil: challenges and opportunities. In: Kurosu, M. (ed.) *HCII/HCI 2013, Part I. LNCS*, vol. 8004, pp. 3–12. Springer, Heidelberg (2013)
12. Boscaroli, C., Silveira, M.S., Prates, R.O., Bim, S.A., Barbosa, S.D.J.: Charting the landscape of HCI education in brazil. In: Kurosu, M. (ed.) *HCI 2014, Part I. LNCS*, vol. 8510, pp. 177–186. Springer, Heidelberg (2014)
13. de Souza, C.S.: *The Semiotic Engineering of Human-Computer Interaction*. The MIT Press, Cambridge (2005)
14. de Souza, C.S., Leitão, C.F., Prates, R.O., Silva, E.J.: The semiotic inspection method. In: VII Brazilian Symposium on Human Factors In Computing Systems (IHC 2006). pp. 148–157. Sociedade Brasileira de Computação (2006)

15. Medeiros, C.B.: Grand research challenges in computer science in brazil. Computer **41**(6), 59–65 (2008)
16. Barbosa, S.D.J., de Souza, C.S.: Are HCI researchers and endangered species in Brazil? ACM Interact. **2011**, 69–71 (2011)
17. Prates, R., Barbosa, S.D.J., da Silveira, M., de Souza, C., Baranauskas, C., Maciel, C., Furtado, E., Anacleto, J., Melo, P., Kujala, T.: HCI community in Brazil- sweet 16! ACM Interact. **20**(6), 80–81 (2013)

# MindMiner: A Mixed-Initiative Interface for Interactive Distance Metric Learning

Xiangmin Fan<sup>1</sup>, Youming Liu<sup>1</sup>, Nan Cao<sup>2</sup>, Jason Hong<sup>3</sup>,  
and Jingtao Wang<sup>1()</sup>

<sup>1</sup> Computer Science and LRDC, University of Pittsburgh, Pittsburgh, PA, USA  
`{xiangmin, youmingliu, jingtaow}@cs.pitt.edu`

<sup>2</sup> IBM T.J. Watson Research, Yorktown Heights, NY, USA  
`nancao@us.ibm.com`

<sup>3</sup> HCI Institute, Carnegie Mellon University, Pittsburgh, PA, USA  
`jasonh@cs.cmu.edu`

**Abstract.** We present MindMiner, a mixed-initiative interface for capturing subjective similarity measurements via a combination of new interaction techniques and machine learning algorithms. MindMiner collects qualitative, hard to express similarity measurements from users via *active polling with uncertainty* and *example based visual constraint creation*. MindMiner also formulates human prior knowledge into a set of inequalities and learns a quantitative similarity distance metric via *convex optimization*. In a 12-subject peer-review understanding task, we found MindMiner was easy to learn and use, and could capture users' implicit knowledge about writing performance and cluster target entities into groups that match subjects' mental models. We also found that MindMiner's constraint suggestions and uncertainty polling functions could improve both efficiency and the quality of clustering.

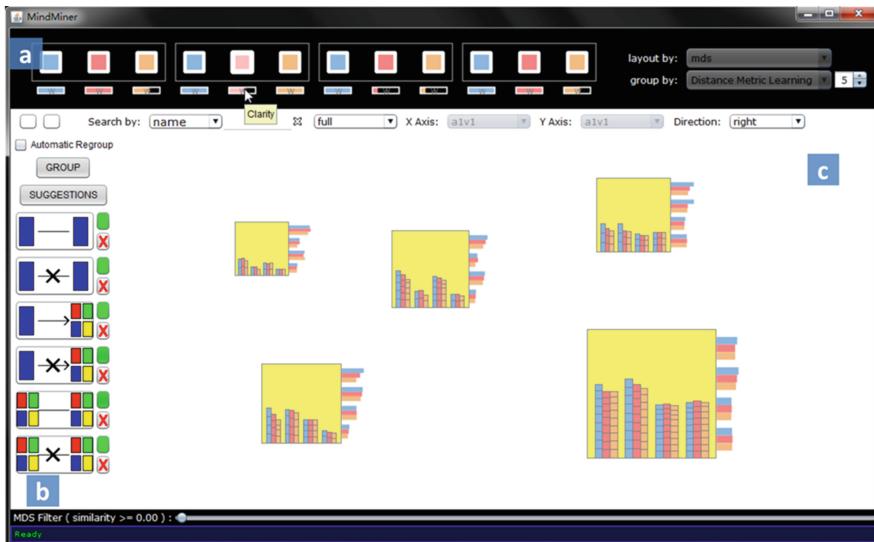
**Keywords:** Mixed-Initiative interface · Clustering · Visualization · Convex optimization · Intelligent user interfaces · Machine learning

## 1 Introduction

Cluster analysis is a common task in exploratory data mining, and involves combining entities with similar properties into groups. Clustering is desirable in that it is unsupervised and can discover the underlying structure of data without *a priori* information. However, most clustering techniques face one key challenge when used in real world applications: clustering algorithms expect a quantitative, deterministic distance function to quantify the similarity between two entities. In most real world problems, such similarity measurements usually require subjective domain knowledge that can be hard for users to explain. For example, a human instructor may easily find that the writing styles of two students are very similar to each other by reviewing their writing samples. However, such perceived similarities may not be reflected accurately in the distance measurement between two corresponding feature vectors.

Previous efforts have been made by researchers to improve the quality of clustering using both algorithmic [8, 27, 28] and user interface [6, 10, 11] approaches. For

example, various semi-supervised clustering algorithms have been proposed by researchers in the machine learning community, either by adapting a similarity measure via user specified constraints or by modifying the process of determining intermediate cluster centers. However, most existing work focuses on *theoretical feasibility*: they assume users can provide sufficient, unambiguous, and consistent information to facilitate clustering before the algorithms start.



**Fig. 1.** The primary UI of MindMiner, showing 23 students in a college-level philosophy class grouped into five clusters based on their performance (accuracy, clarity, and insight) in four writing assignments using six example constraints specified by an instructor. MindMiner consists of three parts: (a) The *Active Polling Panel* allows users to optionally indicate the importance for each measurement. Each colored square box represents one feature ( $4 \text{ assignments} \times 3 \text{ features}$ ). The rectangular bars beneath show real-time updates of the corresponding “weights”; (b) The *Constraints Management Sidebar* displays example-based constraints collected; (c) The *Interactive Visualization Workspace* lets a user see detailed information about entities, create example-based constraints, split and combine groups, examine and refine clustering results and examine personalized groups.

Researchers in HCI and Information Visualization have also explored the use of interactive applications for guided clustering [10, 16, 20, 24]. Some interfaces rely on real time feedback of clustering results to help users choose proper features, samples, and the number of clusters to use. Other systems, such as IVC [10], attempt to provide mechanisms to collect users’ *a priori* knowledge, such as which samples should be in the same group, and which should not. However, most existing interactive clustering systems focus on *conceptual demonstration* and do not address important elements for making such systems practical, such as how to browse, how to manage users’ collected

*a priori* knowledge, and how to achieve better clustering results with more representative constraint examples.

To address these challenges, we created a mixed-initiative interface, MindMiner (Fig. 1), to capture users' subjective similarity measurements. MindMiner makes contributions in both interaction design and machine learning algorithms. MindMiner captures prior knowledge from users through *active polling with uncertainty* and *example based visual constraint creation*. *Active polling with uncertainty* enables users to specify their subjective opinion on the *global* importance of a feature (including the value "not sure") which improves the accuracy and speed of the clustering results. *Example based visual constraint creation* allows to directly express their *a priori* domain knowledge via six types of constraints on the data samples being visualized. The constraint management interface allows users to browse existing examples, investigate the impact of each constraint, and discover conflicting conditions.

MindMiner also provides interface level support that uses *active learning* to provide optional hints as to which examples might be more helpful for clustering. We also report how inequalities are formulated based on the collected *a priori* knowledge and how the inequalities are used in a convex optimization process to extract the "mental model" of entity similarity from users in the form of the Mahalanobis distance metric.

Specifically, this paper makes the following contributions:

- We propose two interaction techniques, *active polling with uncertainty* and *example-based constraints collection*, to collect, visualize, and manage implicit, subjective domain knowledge by scaffolding end-users incrementally. These techniques assume that users' domain knowledge may be *ambiguous* and *inconsistent*.
- We introduce an improved distance metric learning algorithm that takes into account input ambiguity and avoids trivial solutions<sup>1</sup> in existing algorithms.
- We present effective *active learning* heuristics and corresponding interface design to collect pairwise constraints at both entity and group levels. We show in a 12-subject controlled study that our design can significantly enhance the clustering relevance.
- We present an interactive data exploration and visualization system, MindMiner, to help end-users externalize domain knowledge and improve data exploration efficiency via distance metric learning. To our knowledge, this is the first interactive system that provides both algorithm and interface level support for handling *inconsistent*, *ambiguous* domain knowledge via distance metric learning.

## 2 MindMiner in Action

We present a scenario giving an overview of MindMiner. MindMiner was originally designed for computer assisted peer-review and grading scenarios, but can also be used for other interactive clustering tasks.

---

<sup>1</sup> When the number of constraints is small, e.g., less than 20, existing algorithms tend to generate trivial distance metrics that have only one or two non-zero dimensions.

Alice is an instructor for a philosophy course with 23 students. There are four writing assignments, and the essays submitted by students are graded via three features (accuracy, clarity, and insight). The grading is done by herself, the TA, and “double-blind” peer-review by students. Alice feels it is tedious and time consuming to get a clear picture of the overall performance of the whole class. Alice also wants to identify students with similar writing problems so that she can provide customized feedback to them. Alice can use MindMiner to achieve a balance between workload and feedback accuracy.



**Fig. 2.** Knowledge collection interfaces of MindMiner. a: Interface for *active polling with uncertainty*. b: Interface for *example-based constraints collection*.

After logging into MindMiner, Alice retrieves student performance data from a remote server. Alice believes that writing accuracy is the most important factor she cares about and clarity a close second. She is not sure about the importance of insight. Therefore, she uses the *Active Polling Panel* (Fig. 2a) to make a choice for each feature. She chooses “very important” for accuracy, “important” for clarity and “not sure” for insight.

Then Alice teaches MindMiner her subjective judgments on performance similarity of students by labeling some example constraints. Alice reviews detailed information of the students by mousing over the nodes. MindMiner automatically selects the most potentially informative pairs and highlights the suggestions with dashed lines (Fig. 2b). She examines two students involved in a constraint suggestion. After judging that they performed similarly, she drags them together, which creates a must-link constraint between the two students, telling MindMiner that these students should be grouped together. A corresponding symbol for this constraint then appears in the *Constraints Management Sidebar* (Fig. 1b). She later creates a cannot-link between dissimilar students by right clicking and dragging from one to the other. Every time Alice adds a new constraint, the distance metric learning module runs a convex optimization algorithm to derive the optimized solution. The bars in the *Active Polling Panel* (Fig. 1a) show the updated weights of corresponding feature dimensions in real-time.

MindMiner also checks if there are any conflicts caused by new constraints. If so, it gives a warning by highlighting the corresponding constraints in the *Constraints Management Sidebar* using a red background. Alice checks the conflicting constraints

and finds that one of the previous example constraints she created is not correct so she deletes it. Each constraint item in the *Constraints Management Sidebar* is double-linked with corresponding students via mouse hovering, so it is easy for Alice to diagnose the cause when a conflict is reported by MindMiner.

Alice clicks the “group” button located on the top of the *Constraints Sidebar* to see whether the examples provided by her are sufficient for grouping students together in a useful manner. MindMiner applies the updated distance metric using a k-means clustering algorithm, and then displays the resulting groups. Alice then checks the results and finds that the groups are not as good as she expected. She adds a few more constraints and then she checks “automatic regroup”. In this mode, once there is a new constraint, MindMiner’s learning algorithm executes and the system automatically regroups the students based on the most updated distance metric. Alice continues this iterative process by adding new constraints, deleting existing constraints or adjusting importance levels of the features, until she gets satisfactory clustering results.

### 3 Related Work

We have organized related work into three categories: interactive machine learning, clustering interfaces, and semi-supervised clustering algorithms.

#### 3.1 Interactive Machine Learning

Because of the inherent ambiguities in human activities and decision-making processes, many researchers believe that machine learning algorithms can never be perfect in replacing human experts [25]. As an alternative, researchers have investigated mixed-initiative interfaces [15] that keep humans in the loop, providing proper feedback and domain knowledge to machine learning algorithms [1, 2, 7, 12, 18, 21, 26].

For example, CueFlik [12] allows end-users to locate images on the web through a combination of keyword search and iterative example-based concept refinement activities. Although both CueFlik and MindMiner support distance metric learning and active learning, there are major differences between the two systems. First, CueFlik supports one-class information retrieval while MindMiner focuses on multi-group semi-supervised clustering that matches a user’s mental model. Second, in CueFlik, users’ feedback was provided in the form of positive and negative image examples and it is not necessary to handle conflicting examples due to the diversity of online images. In comparison, MindMiner collects both feature uncertainty information and pairwise relationship information from users to formulate a convex optimization problem. MindMiner also provides a dedicated constraint management interface to collect, browse user-specific knowledge and resolve prospective knowledge conflicts, which are common in multi-group clustering tasks.

Apolo [7] is an interactive sense making system intended to recommend new nodes in a large network by letting users specify exemplars for intended groups. Apolo is similar to MindMiner in that both systems accept examples for intended groups from end-users. However, Apolo focuses on suggesting new members for existing groups in

a graph topology, whereas MindMiner aims to generalize examples from end-users, using existing groups to create new groups via distance metric learning. MindMiner also works on an unstructured, unlabeled sample space rather than a graph.

In addition to providing representative examples [7, 12], interactive machine learning also allows end-users to specify their preferences on system output and adapt to these preferences in model parameters [18, 21, 26]. ManiMatrix [18] lets users interactively indicate their preferences on classification results by refining parameters of a confusion matrix. In CAAD [21], users correct classification errors caused by the activity detection algorithm by manually moving documents to the correct category. The input matrix for the Nonnegative Matrix Factorization (NMF) algorithm in CAAD also gets updated by such changes.

### 3.2 Clustering Interfaces

Many interactive tools have been designed to facilitate cluster analysis and exploration by providing real time feedback to parameter and dataset changes. For example, Hierarchical Cluster Explorer (HCE) [24] is an interactive hierarchical clustering system that supports comparison and dynamic querying of clusters. DICON [6] uses a tree-map style, icon-based group visualizations and a combination of k-means clustering and manual grouping to facilitate cluster evaluation and comparison. NodeTrix [14] combines a matrix representation for graphs with traditional node-link graph visualization techniques. Users can select and group nodes to generate an adjacency matrix to visualize cluster patterns in a graph. Many researches also focus on finding clusters in multidimensional data based on parallel coordinate plots (PCPs) [17]. For example, Fua et al. [13] used hierarchical clustering in PCP to detect clusters. Novotny [19] use polygonal area in PCP to represent clusters. However none of these techniques incorporate end-user feedback to improve clustering.

IVC [10] supports incorporating user specified pairwise constraints in clustering. The authors leveraged the PCK-Means algorithm proposed by Basu, Banerjee and Mooney [3] for clustering, and the pairwise constraints were incorporated into the k-means algorithm as a penalty in the cluster assignment state. In contrast to MindMiner, IVC was only tested in simulations and there was no method to manage constraints.

Basu, Fisher et al. [4] implemented a document clustering system that combines user specified constraints and supervised classification. However, there was no constraint collection and management interface in their system.

### 3.3 Semi-supervised Clustering Algorithms

Researchers in machine learning have explored the use of human knowledge in unsupervised clustering [8, 27, 28], i.e. semi-supervised clustering. The users' prior knowledge was leveraged in semi-supervised clustering algorithms by either adapting the similarity measure or modifying corresponding search-rules. The semi-supervised clustering algorithm in MindMiner was inspired by the one proposed by Xing et al. [28].

Three major revisions in Xing’s algorithms were made in MindMiner to generate higher quality results on real-world data. First, we added the support for user specified feature uncertainty and two additional groups of pairwise constraints. Second, we incorporated an active learning algorithm and corresponding heuristics to improve the quality of constraints collected. Third, we added a regularization step to avoid trivial solutions derived from the convex optimization.

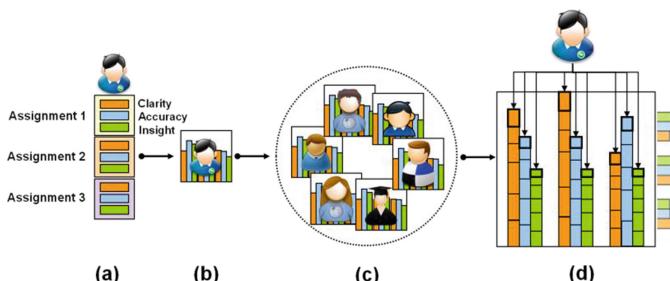
## 4 Design of MindMiner

In the following sections, we discuss these parts in more detail, including the visualization design, the knowledge collection interfaces in MindMiner and the underlying mathematical modeling and the convex optimization algorithm for learning the distance metric respectively.

### 4.1 Visualization Design

We use interactive stacked bar charts in MindMiner to visualize clusters of data with multivariate features. Figure 3 illustrates an example of our design in which a student dataset is visualized. Each student, treated as an entity, is characterized by his/her performances in a writing course along different features, i.e. accuracy, clarity, and insight. These features are defined by the user, and are measured based on the peer-review scores of three writing assignments.

As shown in Fig. 3a, we use different background colors to illustrate different assignments, and use different foreground colors to represent the different features. A student’s feature vector is represented as a bar chart (Fig. 3b) in which the sizes of the bars represent the corresponding review scores. Similarly, we represent a clustered group of students (Fig. 3c) by packing all of the students’ review scores together into a stacked bar chart, categorized by assignments (Fig. 3d). We also represent the averaged student feature scores of each assignment as another grouped bar chart attached to the group. The position of the bar chart, i.e. left, right (default location, Fig. 3d), bottom, and top, can be customized by users. The resulting visualization shows the overall distribution of data while keeping individual details easily visible.



**Fig. 3.** MindMiner visualization design. (a) Feature vector of a student based on three writing assignments and three different features. (b) Student barchart icon. (c) A group of similar students. (d) Stacked bar chart icon for a cluster of students.

## 4.2 Knowledge Collection Interfaces

MindMiner offers two novel knowledge collection techniques, *active polling with uncertainty* and *example-based constraints collection*, to make it easier for end-users to externalize their implicit mental models of entity similarity. We also introduce an active learning [9] heuristic to help users provide similarity examples that are more informative to the follow-up learning algorithms.

**Active Polling with Uncertainty.** MindMiner lets users specify their perceived importance of each feature via Active polling with uncertainty (Fig. 2a). Available choices are – “not important”, “important”, “very important” and “not sure”. This step is optional and the default choice is “not sure”. These choices correspond to different parameter search spaces in the convex optimization stage. As we illustrate later, expressing subjective certainty can reduce the number of examples needed in the next step and improve clustering quality.

**Example-Based Constraints Collection.** MindMiner allows users to specify their knowledge on entity similarity via examples. This approach is supported by a psychology theory [23], which suggests that people represent categories through examples or prototypes. Instead of collecting absolute annotations or labels from end-users, which have been proven by many research findings to be unreliable, especially for subjective domain knowledge, we choose to collect *pairwise* knowledge instead.

End-users can provide three types of constraints to represent their prior knowledge on entity similarity (Table 1): (1) pairwise entity similarity (must-link, cannot-link);

**Table 1.** Symbols and descriptions of the six pairwise constraints supported by MindMiner. Collected constraints are shown in the *Constraints Management Sidebar* (Fig. 1b) (Color figure online).

Symbol	Name	Details
	Must-link	Lets user specify that two entities should be grouped together. Leads to a new entry in equation (2)
	Cannot-link	Lets user specify that two entities should not be grouped together. Leads to a new entry in equation (3).
	Must-belong	Lets user specify that one entity should be included in a specific group. Leads to multiple must-links, and added as multiple entries in equation (2)
	Cannot-belong	Lets user specify that one entity should not be included in a specific group. Leads to multiple cannot-links, and added as multiple entries in equation (3)
	Similar groups	Lets user specify that two existing groups should be put together. Leads to multiple must-links, and added as multiple entries in equation (2)
	Dissimilar groups	Lets user specify that no items in the two existing groups should be put into the other group. Leads to multiple cannot-links, and added as multiple entries in equation (3)

(2) entity-group similarity (must-belong, cannot-belong); (3) pairwise group similarity (similar-groups, dissimilar-groups). The latter two types include new constraints that are never found in existing semi-supervised clustering algorithms. All six constraints can be specified by users in the primary interface via mouse-based direct manipulation operations (Table 1). Constraints created are shown in the *Constraint Management Sidebar* (Fig. 1b). This sidebar allows users to browse, remove, or check the impact of each constraint created. Conflicting constraints are also highlighted in red. In the clustering stage, the top-right corner of each constraint in the *Constraints Management Sidebar* shows whether this constraint is currently satisfied (green means satisfied while red means no). Using visual feedback, rather than directly enforcing them via heuristics, allows end-users to inspect the unsatisfied constraints and refine them if necessary.

Two challenges arise when collecting similarity samples. First, not all constraints are equally useful. A user could provide multiple examples, but it might not improve the convergence speed of the convex optimization algorithm or reliability of the distance metric learning process. Second, investigating the similarity between two entities or groups can be repetitive, tedious, and demanding on short-term memory.

To address these challenges, MindMiner uses an active learning approach to automatically select the entity/group pairs that could be most informative to the follow-up convex optimization algorithm, and then encourages users to specify their similarities. The detailed active learning algorithm and heuristics are described in detail in the next section. Distance Metric Learning Algorithms.

Once MindMiner collects a new piece of information (feature uncertainty or pairwise sample similarity) from users, it converts such information into a set of inequalities, formulates a convex optimization problem [5], and learns a distance metric from user provided similarity information. This distance metric is then used for clustering entities into groups. There are four major steps in this distance metric learning process: (1) constraint conflict detection; (2) inequalities generation; (3) convex optimization; (4) results regularization. We explain details of each step in the rest of this section.

### 4.3 Mathematical Background

An entity in MindMiner is denoted by an  $n$ -dimensional feature vector. For example, entity  $s_i$  is represented by  $(s_{i1}, s_{i2}, \dots, s_{in})$  in which  $n$  is the dimension in the feature space. The similarity measurement  $d(s_i, s_j)$  between entity  $s_i$  and entity  $s_j$  is defined as:

$$d(s_i, s_j) = \sqrt{(s_i - s_j)W(s_i - s_j)^T} \quad (1)$$

Here  $W$  is an  $n \times n$  distance metric matrix. Letting  $W = I$  leads to Euclidean distance. In MindMiner, we restrict  $W$  to be diagonal for efficiency concerns, the same framework can be used to learn a complete  $W$  with sufficient user examples. Determining

each non-zero element in the diagonal  $W$  corresponds to learning a metric in which the different measurement features are given different “weights”. Therefore, our goal here is to find  $W$  (weights vector) which best respects the information collected via the active polling process and interactive constraint creation process.

#### 4.4 Constraint Conflict Detection

The information collected with *active polling with uncertainty* is used to define the lower and upper bound of the associated weight for each feature in the follow-up optimization process. The choice “Very important” corresponds to a weight of 1 (highest), “not important” corresponds to a weight of 0 (lowest), the weights of “important” features are set to be in a range of [0.6, 1] while “not sure” features are set to be within [0, 1]. In the end, we get a set of ranges for the weights of all features:

$$\text{WeightBounds} (WB) = \{[w_{1lb}, w_{1ub}], \dots, [w_{nlb}, w_{nub}]\}$$

As shown in Table 1, depending on the constraint type, each constraint collected will be converted to one or multiple pairwise relationships and a Boolean flag. For must-link and cannot-link, the corresponding list only contains one pair, with a Boolean flag indicating the similarity relationship (true for similar and false for dissimilar) between the entities involved in the pair. For other types of constraints, they are first converted to multiple pairwise constraints such as must-links or cannot-links. Then these must-links or cannot-links are added to the pairs list of the corresponding constraint.

```

C_i \in C do
    if the similarity flags of c and  $C_i$  are different then
        //conflicts may exist;
        iterate through the pairs lists of c and  $C_i$  to see if there is a
        common pair. If yes, mark  $C_i$  as a conflicting constraint.
    end
end

```

**Algorithm 1.** Constraint conflict detection.

By using this list based constraint representation, Algorithm 1 presents pseudo code to detect prospective conflicts in the constraints provided by end-users. If a constraint conflict is detected, corresponding constraints in the *Constraints Management Sidebar* (Fig. 1b) will turn red. Also, hovering over a conflicting constraint will highlight the remaining constraint(s) in conflict, as well as the corresponding entities and groups.

#### 4.5 Active Learning Heuristic

As noted earlier, not all user-specified examples are equally helpful in improving the results from convex optimization. Some examples could be repetitive and would not

justify the time spend by users to specify them or the extra computer-time added to the optimization process. To address this dilemma, we adopted concept of active learning, which allows MindMiner to identify and suggest ambiguous entity relationships that are most informative in improving the quality of distance metric learning. For example, suppose that an entity  $S_1$  is currently in cluster A and  $S_1$  is on the boundary of A and B while student  $S_2$  is the best exemplar of A; then the constraint suggestion  $\langle S_1, S_2 \rangle$  would be posed to end-users asking for whether they are similar or not. We designed a three-step active learning heuristic listed below to recommend informative constraint samples to end users. This active learning heuristic will be executed every time when a new constraint is provided by users. The informative entity pairs discovered via active learning are marked with dashed lines in the main interface.

- Within each cluster  $c$ , find the entity with minimum distance to the center of  $c$  as the exemplar for  $c$ .
- For each entity  $s$ , calculate the difference  $d$  between the distances from  $s$  to the nearest two clusters  $c_1$  and  $c_2$ . If  $d$  is less than a threshold, we mark the entity  $s$  as ambiguous.
- For each entity  $s'$  that was marked as ambiguous, create a constraint suggestion between  $s'$  and the exemplar of cluster it currently belongs to.

#### 4.6 Inequality Generation

We also keep two global sets:  $S$ , which is a set of pairs of entities to be “similar” and  $D$ , which is a set of pairs of entities to be “dissimilar”. All the similar pairs are added to  $S$  while all the dissimilar pairs are added to  $D$  during the interactive constraint creation process.

After the constraint conflict detection step, we convert the user knowledge collected through *active polling with uncertainty* and *example-based constraints collection* to Weight Bounds which are a set of weight ranges for all features, and  $S$  and  $D$  which are sets of pairs of similar/dissimilar entities.

A straightforward way of defining a criterion for the meaningful distance metric is to demand that pairs of entities in  $S$  have small squared distance between them (Eq. 2). However, this is trivially solved with  $W = 0$  and is not informative. Our approach was primarily inspired by the method proposed by Xing et al. [28]. To avoid the above mentioned trivial solution, we add a new inequality constraint (Eq. 3) to ensure it takes dissimilar entities apart. In this framework, we transform the problem of learning meaningful distance metrics to a convex optimization problem:

$$\min_w \sum_{(s_i, s_j) \in S} d^2(s_i, s_j) \quad (2)$$

s.t.

$$\sum_{(s_i, s_j) \in D} d(s_i, s_j) \geq 1 \quad (3)$$

For each

$$w_k : w_k \geq 0 (1 \leq k \leq n) \quad (4)$$

Each sum item in Eq. 2 corresponds to a positive constraint collected, while each sum item in Eq. 3 corresponds to a negative constraint collected (Table 1).

It can be proven that the optimization problem defined by Eqs. 2–4 is convex, and the distance metric  $W_{raw}$  can be solved by efficient, local-minima-free optimization algorithms.

Unfortunately, according to our early experiences on real world data, it is not desirable to use  $W_{raw}$  as the distance metric for the follow-up clustering tasks. According to our observations, when the number of constraints is very small, especially at the beginning of a task, convex optimization usually lead to a sparse distance metric where most values in the distance metric are close to zeros, i.e. only minimal features, e.g., 1 or 2 features, are taken into account in similarity measurement, implying a trivial solution that does not represent the real-world situation. We use an extra result regularization step and leverage the information collected in the *active polling with uncertainty* step to generate more meaningful distance metric that could be a better representation of a user's mental model.

## 4.7 Result Regularization

In order to make distance metrics respect both feature uncertainty information and the constraints collected by MindMiner, we regularize  $W_{raw}$  by using *Weight Bounds (WB)*. Detailed steps are described in Algorithm 2.

After finishing the result regularization step, we get a  $W$  that conforms to all the prior knowledge we collected from end-users. We apply  $W$  to the distance metric function and get the relevant distance metric. Then the distance metric  $W$  is used in k-means clustering algorithm to generate meaningful clusters.

```

W_{raw}; and the “lower bound” ( $W_{i_{lb}}$ ) and “upper
      bound” ( $W_{i_{ub}}$ )
output: the regularized weight  $W_i$  ( $1 \leq i \leq n$ )
Iterate through  $W_{raw}$  and find the maximum and minimum values
 $W_{raw\_max}, W_{raw\_min}$ 
for  $i$  from 1 to  $n$  do
  if  $W_{i_{lb}} = W_{i_{ub}} = 1$  then
    | set  $W_i = 1$ 
  else if  $W_{i_{lb}} = W_{i_{ub}} = 0$  then
    | set  $W_i = 0$ 
  else
    | set  $W_i = W_{i_{lb}} + (W_{i_{ub}} - W_{i_{lb}}) \cdot \frac{W_{raw_i} - W_{raw\_min}}{W_{raw\_max} - W_{raw\_min}}$ 
end

```

**Algorithm 2.** Result Regularization.

## 4.8 Implementation

MindMiner is written in Java 1.6. The convex optimization algorithm is implemented in Matlab. We use Matlab Builder JA to generate a Java wrapper (i.e. a jar file) around the actual Matlab programs. MindMiner can run as a desktop application or a web application via Java Web Start.

## 5 Evaluation

We conducted a 12-subject user study to understand the performance and usability of MindMiner. We had three basic goals for this study. One was to figure out whether or not the ideas behind MindMiner are easy to understand and if the current MindMiner implementation is easy to learn. The second goal was to evaluate the overall performance of MindMiner in capturing the similarity measurements in users' minds. The third goal was to verify the efficacy of each new component designed (i.e. *active learning heuristics*, *example based visual constraint creation*, and *active polling with uncertainty*). The data loaded in MindMiner in this study was anonymized real world data from a 23 student philosophy course in a local university with permission from the internal review board (IRB) and the instructor.

### 5.1 Experimental Design

The study consisted of five parts:

**Overview.** We first gave participants a brief introduction and a live demo of MindMiner. We explained each task to them, and answered their questions. After the introduction, we let the participants explore the interface freely until they stated explicitly that they were ready to start the follow-up tasks.

**Clustering and Active Learning.** We used a within-subjects design in this session. There were two similar tasks: task 1 was clustering the students into four groups based on their performance in the first assignment; task 2 was the same as the previous task except that users were to only consider the “accuracy” features of the assignments. There were two conditions in this section: (A) providing constraint suggestions via active learning; (B) without active learning. Six participants performed task 1 with condition A and task 2 with condition B. The other six performed task 1 with condition B and task 2 with condition A. The order of the two tasks was counter-balanced. Each participant could provide up to ten example-based pairwise constraints (both positive examples and negative examples) for each task. The active polling with uncertainty feature was disabled in both conditions. We collected each participant's task completion time for each condition and the distance metrics derived by the learning algorithm.

**Active Polling with Uncertainty.** We used a between-subjects design in this session with two conditions: the constraints & active polling condition and the constraints-only condition. The active learning feature was enabled in both conditions. The task required users to find five students with similar performances to one student named “Indrek”. We told the participants that the accuracy and clarity features of the

first two assignments were very important to consider and asked them to define the importance of other features themselves. We hypothesized that given meaningful clustering results, one can find similar students easily just by going over each student in the target's group. Otherwise, if the clustering was not good, the participants would have to view groups besides the target's group to find similar students.

**Free Exploration.** In this session, the participants were asked to group the students into three categories based on their own grouping criteria. Users were encouraged to think aloud and even write down their rules on a piece of paper. They were also encouraged to explore MindMiner as long as they wanted.

**Qualitative Feedback.** After participants completed all the tasks, they were asked to complete a questionnaire and describe their general feeling towards our system.

## 5.2 Participants and Apparatus

We recruited 12 participants (5 female) between 22 and 51 years of age (mean = 27) from a local university. Two were instructors from physics department and psychology department respectively. The other ten were graduate students who have teaching experience. Each study lasted for around 60 min (up to 90 min maximum), and each participant was given a \$10 gift card for the time.

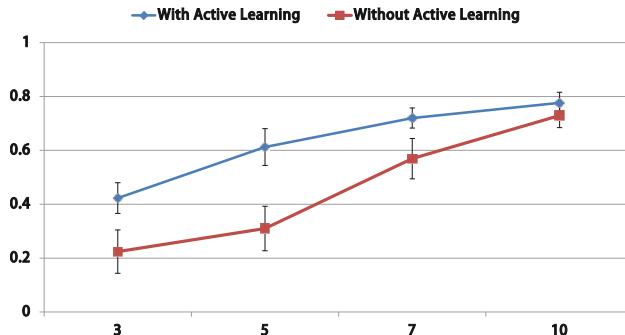
A Lenovo ThinkPad T530 laptop computer with Intel Core i5-3210 CPU, 4 GB RAM, running Windows 7 was used. An external NEC 23 inch LCD monitor with a resolution of 1920\*1080 was attached to the laptop to run MindMiner.

## 5.3 Evaluation Results

**Clustering and Active Learning.** The average task completion time in the “with active learning” condition is significantly shorter than that of the “without active learning” condition (266.4 s vs. 357.4 s,  $F_{1, 11} = 13.403$ ,  $p < 0.01$ ). We observed that with active learning suggestions enabled, participants tended to compare the students involved, instead of randomly picking several students to compare. MindMiner suggestions gave them clear “targets” to inspect; otherwise, they would look for “targets” themselves, which usually leads to more time. Furthermore, when there were no suggestions, participants had to compare multiple students, which required having to remember many students’ scores. In comparison when they had system suggestions, they only need to compare two students.

To evaluate the quality of distance metrics learned in the two conditions, we defined our “gold standard” to be a weight vector where the weights of predefined important features are 1 s, and the weights of other features are 0 s. We used cosine similarity between the standard weight vector and the weight vector learned from our algorithm to measure the quality of distance metric learned (Fig. 4).

Analysis of variance revealed that there was a significant difference ( $F_{1, 11} = 7.42$ ,  $p < 0.05$ ) in the quality of the distance metric learned. We found that there was a significant main effect ( $F_{3, 9} = 19.30$ ,  $p < 0.05$ ) in quality among different numbers of

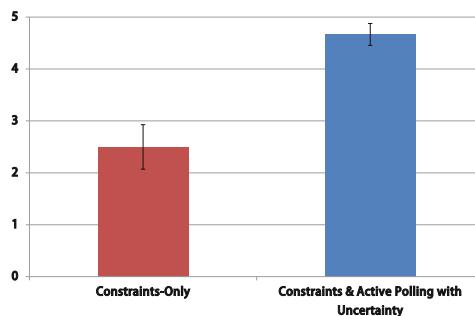


**Fig. 4.** Average cosine similarities between “gold standard” and distance metrics learned by different numbers of constraints (the higher the better).

constraints collected. Pairwise mean comparison showed that more constraints led to significantly better quality distance metrics. With the same number of constraints, the quality of distance metrics learned with active learning was significantly higher than that without active learning for all four numbers of constraints in Fig. 4.

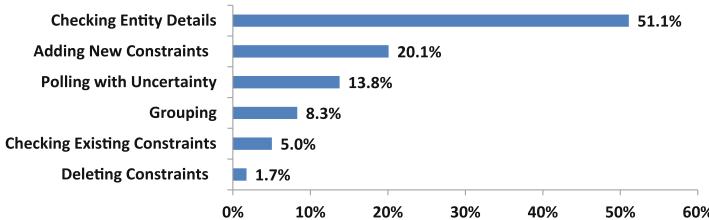
**Active Polling with Uncertainty.** When active polling with uncertainty was enabled, the average completion time was 252.7 s ( $\sigma = 19.6$ ). When disabled, the average completion time was 304.8 s ( $\sigma = 43.1$ ). However, the difference was not statistically significant ( $p = 0.297$ ).

The active polling with uncertainty condition also led to significantly more similar students discovered (4.67 vs. 2.50,  $p < 0.001$ ) than the condition without active polling (Fig. 5). This finding showed that active polling with uncertainty could also facilitate users by helping them to learning process to derive more relevant entities.



**Fig. 5.** Average number of similar students discovered by condition (the more the better).

**Free Exploration.** A total of 458 interaction activities were recorded in the free exploration session (Fig. 6). Examining details panels was the most frequent activity (51.1 %), followed by adding constraints (20.1 %). Of all the 92 constraints added by



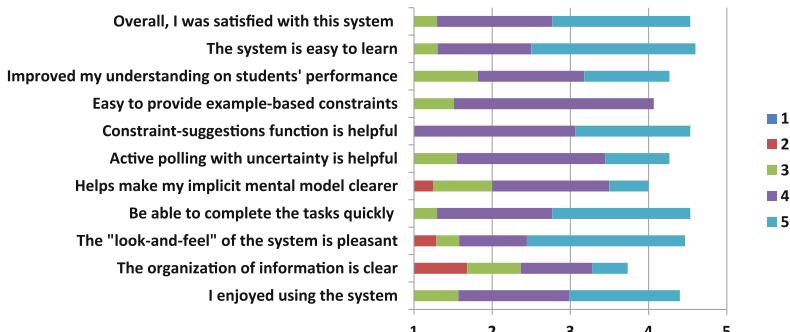
**Fig. 6.** Activity distribution of participants.

participants in this session, 74 (80.4 %) were from system suggestions. Other frequent activities included using the active polling with uncertainty feature (13.8 %), grouping entities (8.3 %), checking existing constraints (5.0 %), and deleting constraints (1.7 %). Among all the 8 constraint deletions, 6 were unsatisfied inappropriate constraints and 2 were constraint conflicts.

We observed that participants tended to add more positive examples (must-link, must-belong, and similar-groups) than negative examples (cannot-link, cannot-belong, and dissimilar-groups) (78.6 % vs. 21.4 %) when the active learning feature was disabled. Participants tend to not provide negative examples even when they were confident that two entities were very different; when the active learning feature was enabled, the ratio of negative examples almost doubled (40.8 %) and the difference was statistically significant. This observation indicated that the current active learning interface and heuristics in MindMiner can increase users' awareness and contribution to negative examples.

Although participants were encouraged to take a look at the suggested entity relationships first before searching for their own examples in the active learning condition, some subjects chose not to do so. When asked for why, the reasons were either that they didn't completely trust the computer or that they simply enjoyed the feeling of finding examples from scratch. In either case, the active learning suggestions provided hints for reducing their example finding efforts.

**Subjective Feedback.** Overall, participants reported positive experiences with MindMiner. Participants felt that the system improved their understanding of students' performance through peer-review data (Fig. 7).



**Fig. 7.** Subjective ratings [22] on a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree).

## 6 Conclusion

We presented MindMiner, a mixed-initiative interface to capture domain experts' subjective similarity measurements via a combination of new interaction techniques and machine learning algorithms. MindMiner collects qualitative, hard to express similarity measurements from users via *active polling with uncertainty*, *example based visual constraint creation* and *active learning*. MindMiner also formulates human prior knowledge into a set of inequalities and learns a quantitative similarity distance metric via convex optimization. In a 12-subject user study, we found that (1) MindMiner can capture the implicit similarity measurement from users via *examples collection* and *uncertainty polling*; (2) *active learning* could significantly improve the quality of distance metric learning when the same numbers of constraints were collected; (3) the *active polling with uncertainty* method could improve the task completion speed and result quality.

## References

1. Amershi, S., Fogarty, J., Kapoor, A., Tan, D.: Effective end-user interaction with machine learning. In: Proceedings of the 25th AAAI Conference on Artificial Intelligence, pp. 1529–1532. AAAI, Menlo Park (2011)
2. Amershi, S., Fogarty, J., Kapoor, A., Tan, D.: Overview based example selection in end user interactive concept learning. In: Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology, pp. 247–256. ACM, New York (2009)
3. Basu, S., Banerjee, A., Mooney, R.J.: Active semi-supervision for pairwise constrained clustering. In: Proceedings of the 2004 SIAM International Conference on Data Mining, vol. 4, pp. 333–344. SIAM, Philadelphia (2004)
4. Basu, S., Fisher, D., Drucker, S.M., Lu, H.: Assisting users with clustering tasks by combining metric learning and classification. In: Proceedings of the 24th AAAI Conference on Artificial Intelligence, pp. 394–400. AAAI, Menlo Park (2010)
5. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
6. Cao, N., Gotz, D., Sun, J., Qu, H.: DICON: interactive visual analysis of multidimensional clusters. IEEE Trans. Vis. Comput. Graph. **17**(12), 2581–2590 (2011). IEEE Press, New York
7. Chau, D., Kittur, A., Hong, J., Faloutsos, C.: Apolo: making sense of large network data by combining rich user interaction and machine learning. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 167–176 ACM, New York (2011)
8. Cohn, D., Caruana, R., McCallum, A.: Semi-supervised clustering with user feedback. Constrained Clust. Adv. Algorithms Theor. Appl. **4**(1), 17–32 (2003)
9. Cohn, D., Atlas, L., Ladner, R.: Improving generalization with active learning. Mach. Learn. **15**(2), 201–221 (1994)
10. DesJardins, M., MacGlashan, J., Ferraioli, J.: Interactive visual clustering. In: Proceedings of the 12th International Conference on Intelligent User Interfaces, pp. 361–364. ACM, New York (2007)

11. Dy, J., Brodley, C.: Visualization and interactive feature selection for unsupervised data. In: Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 360–364. ACM, New York (2000)
12. Fogarty, J., Tan, D., Kapoor, A., Winder, S.: CueFlik: interactive concept learning in image search. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 29–38. ACM, New York (2008)
13. Fua, Y.H., Ward, M.O., Rundensteiner, E.A.: Hierarchical parallel coordinates for exploration of large datasets. In: Proceedings of the conference on Visualization 1999: celebrating ten years, pp. 43–50. IEEE Computer Society Press, New York (1999)
14. Henry, N., Fekete, J., McGuffin, M.J.: NodeTrix: a hybrid visualization of social networks. *IEEE Trans. Vis. Comput. Graph.* **13**(6), 1302–1309 (2007). IEEE Press, New York
15. Horvitz, E.: Principles of mixed-initiative user interfaces. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 159–166. ACM, New York (1999)
16. Huang, Y., Mitchell, T.: Exploring hierarchical user feedback in email clustering. In: EMAIL 2008: Proceedings of the Workshop on Enhanced Messaging-AAAI, pp. 36–41. AAAI, Menlo Park (2008)
17. Inselberg, A., Dimsdale, B.: Parallel coordinates: a tool for visualizing multi-dimensional geometry. In: Proceedings of the 1st Conference on Visualization, pp. 23–26. IEEE Press, Washington (1990)
18. Kapoor, A., Lee, B., Tan, D., Horvitz, E.: Interactive optimization for steering machine classification. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1343–1352. ACM, New York (2010)
19. Novotny, M.: Visually effective information visualization of large data. In: Proceedings of the 8th Central European Seminar on Computer Graphics, pp. 41–48 (2004)
20. Peter, A., Shneiderman, B.: Integrating statistics and visualization: case studies of gaining clarity during exploratory data analysis. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 265–274. ACM, New York (2008)
21. Rattenbury, T., Canny, J.: CAAD: an automatic task support system. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 687–696. ACM, New York (2007)
22. Robertson, J.: Stats: we're doing it wrong. <http://cacm.acm.org/blogs/blog-cacm/107125-stats-were-doing-it-wrong/>
23. Rosch, E., Mervis, C.: Family resemblances: studies in the internal structure of categories. *Cogn. Psychol.* **7**(4), 573–605 (1975). Elsevier, Amsterdam
24. Seo, J., Shneiderman, B.: Interactively exploring hierarchical clustering results [gene identification]. *Computer* **35**(7), 80–86 (2002). IEEE, New York
25. Shneiderman, B., Maes, P.: Direct Manipulation vs. Interface Agents. *Interactions* **4**(6), 42–61 (1997). ACM, New York
26. Talbot, J., Lee, B., Kapoor, A., Tan, D.: EnsembleMatrix: interactive visualization to support machine learning with multiple classifiers. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1283–1292. ACM, New York (2009)
27. Wagstaff, K., Cardie, C., Rogers, S., Schrödl, S.: Constrained K-Means clustering with background knowledge. In: ICML, vol. 1, pp. 577–584 (2001)
28. Xing, E., Ng, A., Jordan, M., Russell, S.: Distance metric learning with application to clustering with side-information. In: Advances in Neural Information Processing Systems, pp. 505–512 (2002)

# Author Index

- Abascal, Julio [I-1](#)  
Abdelnour-Nocera, José [IV-630, IV-659](#)  
Abdelrahman, Yomna [III-376](#)  
Abdollah, Norfarhana [IV-628](#)  
Abouelsaadat, Wael [I-316](#)  
Adagunodo, Emmanuel Rotimi [II-479](#)  
Aedo, Ignacio [I-38](#)  
Afonso, Ana Paula [IV-327](#)  
Ahm, Simon [I-237](#)  
Ahmad, Muneeb Imtiaz [I-436](#)  
Ahmed, Naveed [II-522](#)  
Ahn, Hyeonjung [IV-594](#)  
Akamatsu, Shigenori [IV-569](#)  
Alechina, Natasha [II-265](#)  
Alexander, Jason [I-384, II-315, II-349](#)  
Al-Megren, Shiroq [IV-156](#)  
Alshammari, Mohammad [II-584](#)  
Alt, Florian [IV-300](#)  
AlTarawneh, Ragaad [IV-495](#)  
Álvarez Márquez, Jesús Omar [II-20](#)  
Amrhein, Karsten [III-1](#)  
Anacleto, Junia [III-578, IV-433, IV-647](#)  
Anane, Rachid [II-584](#)  
André, Elisabeth [IV-657](#)  
Ang, Chee Siang [III-367](#)  
Anthony, Lisa [IV-1](#)  
Antunes, Pedro [III-552](#)  
Appriou, Aurélien [IV-354](#)  
Ariza, Oscar [III-259](#)  
Arriaga, Rosa I. [III-53](#)  
Arrue, Myriam [I-1](#)  
Asari, Yusuke [III-296](#)  
Aslan, İlhan [IV-621](#)  
Atkinson, Sarah [II-265](#)  
Aylett, Matthew P. [IV-473](#)  
  
Bader, Moritz [II-428](#)  
Bader, Patrick [IV-282](#)  
Bagg, Tobias [III-402](#)  
Bailly, Gilles [III-221, IV-55](#)  
Bakker, Saskia [IV-456](#)  
Balaam, Madeline [II-531](#)  
Balata, Jan [I-89](#)  
Banchs, Rafael E. [IV-510](#)  
  
Bannon, Liam [IV-667](#)  
Barboni, Eric [II-211](#)  
Barbosa, Simone Diniz Junqueira [II-592, IV-637, IV-667](#)  
Bardill, Andy [IV-409](#)  
Barricelli, Barbara Rita [IV-659](#)  
Barthel, Henning [III-1](#)  
Bartneck, Christoph [I-263](#)  
Bauer, Jens [I-47](#)  
Baumgartner, Axel [II-331](#)  
Bayarri, Clara [I-445](#)  
Begany, Grace [II-249](#)  
Belk, Marios [I-523](#)  
Bellino, Alessio [III-534](#)  
Bellucci, Andrea [I-38](#)  
Bergamo, Marília Lyra [IV-482](#)  
Bernhaupt, Regina [II-412, IV-661](#)  
Bertelsen, Olav W. [IV-667](#)  
Biegler, Stefan [I-219](#)  
Bim, Sílvia Amélia [II-592](#)  
Bimamisa, David [IV-486](#)  
Bird, Jon [IV-264](#)  
Blake, Andrew [III-570](#)  
Bødker, Susanne [IV-667](#)  
Boll, Susanne [IV-582](#)  
Borchers, Jan [I-289](#)  
Boring, Sebastian [III-455](#)  
Bornstein, Jens [I-80](#)  
Boscaroli, Clodis [II-592](#)  
Bourdôt, Patrick [IV-148](#)  
Brauner, Philipp [I-453](#)  
Brereton, Margot [II-150](#)  
Brewster, Stephen [IV-478, IV-611](#)  
Briggs, Pam [II-565](#)  
Bröhl, Christina [IV-514](#)  
Brown, Quincy [IV-1](#)  
Broy, Nora [IV-300](#)  
Bruder, Gerd [III-259](#)  
Bruun, Anders [I-237, II-159](#)  
Bryan-Kinns, Nick [II-47](#)  
Buchanan, George [IV-578](#)  
Buchner, Roland [IV-140](#)  
Bueno, Andre [III-578](#)  
Bullen, Andrew [I-210](#)

- Bulling, Andreas I-384, II-315  
 Burmistrov, Ivan II-106  
 Burnett, Daniel IV-72, IV-586  
 Burstyn, Jesse I-332  
 Burtner, Russ II-463  
 Buschek, Daniel II-428  
 Bützler, Jennifer IV-514
- Cain, Rebecca IV-518  
 Caldeira, Miguel IV-335  
 Calderon, Roberto IV-647  
 Calvary, Gaëlle IV-123  
 Camara, Fatoumata IV-123  
 Campos, Pedro IV-659  
 Cao, Nan II-611  
 Cao, Xiang II-436  
 Cappelletti, Alessandro II-73  
 Casiez, Géry IV-231  
 Castet, Julien I-472  
 Catala, Alejandro II-195  
 Cava, Ricardo II-211  
 Centieiro, Pedro III-341  
 Chaves, Thiago II-395  
 Chen, Fang I-550  
 Cheong, Kakit III-493  
 Chia, Dan IV-510  
 Chong, Ming Ki III-367  
 Clemmensen, Torkil IV-630  
 Coelho, José I-110, I-129  
 Côgo, Filipe Roseiro IV-87  
 Colley, Ashley I-324, IV-363  
 Coninx, Karin II-368, IV-264  
 Constantinides, Argyris I-523  
 Convertino, Gregorio IV-673  
 Cooperstock, Jeremy R. II-1  
 Cortés-Rico, Laura III-518  
 Coulton, Paul IV-72, IV-586  
 Coutrix, Céline I-349  
 Csikszentmihályi, Chris IV-630  
 Cubaud, Pierre I-531  
 Cummings, Danielle I-263  
 Cutrell, Edward II-505
- da Rosa, Isaías Barreto II-479  
 Daiber, Florian III-259  
 Dalsgaard, Peter III-596  
 Darzentas, Jenny IV-665, IV-669  
 Dawson, Carolyn IV-518  
 Day, Jonathan IV-578
- de la Rivière, Jean-Baptiste I-472  
 De Luca, Alexander II-428  
 de Oliveira, Elisa Leo IV-590  
 de Souza, Clarisse S. I-201, IV-667  
 de Souza, Marcella Leandro Costa IV-482  
 Degraen, Donald II-368  
 Dennis, Matt IV-677  
 Denzinger, Jochen IV-640  
 Dermody, Fiona IV-499  
 Desnos, Antoine II-412  
 Dickinson, Thomas IV-473  
 Dillenbourg, Pierre II-81  
 Dimitrokali, Elisavet IV-518  
 Dingler, Tilman III-402  
 Dokas, Ioannis IV-514  
 Donovan, Jared II-134  
 Duarte, Carlos I-110, I-129  
 Duarte, Emanuel Felipe IV-87  
 Dugan, Casey II-38, III-418, IV-671  
 Dunphy, Paul II-565  
 Dünser, Andreas I-263  
 Duysburgh, Pieter II-292
- Eagan, James III-221  
 Ebert, Achim I-47, III-1, IV-495, IV-675  
 Eduardo Dias, A. III-341  
 Eduardo Pérez, J. I-1  
 Edwards, Alistair D.N. I-72, I-147  
 Egger, Sebastian IV-626  
 Ehlers, Jan III-526  
 El-Shimy, Dalia II-1  
 Endert, Alex II-463  
 Engelbrecht, Klaus-Peter IV-633  
 Englert, Frank IV-537  
 Erazo, Orlando III-552  
 Erickson, Thomas II-38  
 Ervasti, Mari II-455
- Fahssi, Racim IV-192  
 Fan, Xiangmin II-611  
 Farren, Margaret IV-499  
 Farrow, Elaine IV-473  
 Feary, Mike IV-663  
 Fechner, Thore II-387  
 Feijns, Loe III-45  
 Fellion, Nicholas I-332  
 Fels, Sidney IV-433, IV-642, IV-647  
 Férey, Nicolas IV-148  
 Ferreira, Alfredo III-622

- Ferreira, Lidia Silva [IV-482](#)  
 Ferreira, Vinicius [III-578](#)  
 Fetter, Mirko [IV-486](#)  
 Feuerstack, Sebastian [IV-105](#)  
 Fickel, Alexander [IV-469](#)  
 Fields, Bob [IV-409](#)  
 Figueiredo, Lucas S. [II-395](#)  
 Filgueiras, Lucia Vilela Leite [I-20](#)  
 Fitzpatrick, Geraldine [II-195](#), [IV-624](#)  
 Flores, Luciano Vargas [IV-590](#)  
 Forbrig, Peter [IV-661](#)  
 Forrester, Ian [IV-72](#), [IV-586](#)  
 Fortmann, Jutta [IV-582](#)  
 Frauenberger, C. [IV-624](#)  
 Freeman, Euan [IV-478](#), [IV-611](#)  
 Freitas, Carla [II-211](#)  
 Frey, Jérémie [I-472](#), [IV-354](#), [IV-381](#)  
 Frishberg, Nancy [IV-673](#)  
 Fröhlich, Peter [IV-626](#)  
 Fuchsberger, Verena [III-203](#), [IV-621](#),  
[IV-659](#)  
 Füchsel, Silke [I-445](#)  
 Funk, Mathias [III-45](#)  
 Furci, Ferdinando [I-165](#)
- Gacem, Hind [III-221](#)  
 Gad-El-Hak, Chalid [IV-448](#)  
 Gallagher, Blaithin [IV-669](#)  
 Ganhör, Roman [IV-448](#), [IV-624](#)  
 Garcia-Sanjuan, Fernando [II-195](#)  
 Gärtner, Magdalena [II-331](#)  
 Gasparini, Isabela [II-592](#)  
 Geerts, David [IV-417](#)  
 Geiger, Chris [III-71](#)  
 Geiselhart, Florian [III-526](#)  
 Gellersen, Hans [I-384](#), [II-315](#), [II-349](#)  
 Germanakos, Panagiotis [I-523](#)  
 Gervais, Renaud [IV-381](#)  
 Geyer, Werner [II-38](#)  
 Ghazali, Masitah [IV-628](#)  
 Ghosh, Sanjay [IV-528](#)  
 Giuliani, Manuel [IV-621](#)  
 Giunchiglia, Fausto [III-149](#)  
 Goguey, Alix [IV-231](#)  
 Gonçalves, Frederica [IV-659](#)  
 Gonçalves, Tiago [IV-327](#)  
 Gosset, Phil [III-410](#)  
 Gradinari, Adrian [IV-72](#), [IV-586](#)  
 Graeff, Delphine [I-472](#)  
 Grau, Yves [III-402](#)
- Grechenig, Thomas [I-219](#), [III-333](#)  
 Greenberg, Saul [II-436](#)  
 Greis, Miriam [IV-256](#)  
 Gross, Tom [IV-640](#)  
 Grubert, Jens [IV-523](#)  
 Grünbauern, Martin Gielsgaard [III-185](#)  
 Grünloh, Christiane [I-298](#)  
 Gu, Jiawei [II-436](#)  
 Gugenheimer, Jan [III-350](#)  
 Guiard, Yves [IV-55](#)  
 Güldenpennig, F. [IV-624](#)  
 Gulliksen, Jan [I-298](#), [I-418](#), [IV-637](#)  
 Gumudavelly, Vijay [IV-213](#)  
 Gunes, Hatice [II-47](#)
- Hachet, Martin [IV-354](#), [IV-381](#)  
 Häikiö, Juha [II-455](#)  
 Häkkilä, Jonna [I-324](#), [III-384](#), [IV-363](#)  
 Halldorsdottir, Gyda [II-98](#)  
 Hamdi, Hamidreza [I-340](#)  
 Hamid, Nazatul Naquiah Abd [I-72](#)  
 Hámorník, Balázs Péter [IV-461](#)  
 Hanna, Julian [IV-335](#)  
 Hanson, Vicki [IV-677](#)  
 Harkke, Ville [II-300](#)  
 Harms, Johannes [I-219](#), [III-333](#)  
 Hartmann, Gerhard [I-298](#)  
 Hassib, Mariam [IV-300](#)  
 Hautasaari, Ari [I-183](#)  
 Hautopp, Heidi [III-132](#)  
 Hayashi, Yuki [IV-319](#)  
 He, Yun [I-324](#)  
 Heinicke, Antje [IV-514](#)  
 Heintz, Matthias [III-501](#)  
 Hendley, Robert J. [II-584](#)  
 Henze, Niels [III-402](#), [IV-256](#), [IV-282](#)  
 Hercegfi, Károly [IV-461](#)  
 Herd, Kate [IV-409](#)  
 Heslop, Philip [II-531](#)  
 Hespanhol, Luke [III-596](#)  
 Hess, Steffen [IV-675](#)  
 Heuten, Wilko [IV-582](#)  
 Hinrichs, Klaus [IV-425](#)  
 Hirose, Michitaka [III-80](#)  
 Hiyama, Atsushi [III-80](#)  
 Hodges, Steve [II-565](#)  
 Hödl, O. [IV-624](#)  
 Holden, Amey [IV-156](#)  
 Hong, Jason [II-611](#)  
 Hook, Jonathan [IV-156](#)

- Hoonhout, Jettie **IV-673**  
 Hornbæk, Kasper **III-455**  
 Hörold, Stephan **IV-537**  
 Hu, Jun **III-45**  
 Huckauf, Anke **III-526**  
 Huffaker, David **III-116**  
 Huizenga, Janine **I-210**  
 Huldtgren, Alina **III-71**  
 Humayoun, Shah Rukh **IV-495, IV-675**  
 Huot, Stéphane **IV-148**  
 Hurtienne, Jörn **IV-248**  
 Hußmann, Heinrich **III-614**  
 Hvannberg, Ebba Thora **II-98**
- Ifukube, Tohru **III-80**  
 Ituka, Takakuni **IV-533**  
 Iivari, Netta **III-9**  
 Ikeda, Mio **IV-248**  
 Inamura, Noriko **III-80**  
 Inoue, Yosho **IV-533, IV-563**  
 Ioannou, Andri **II-55**  
 Ishii, Ryo **IV-319**  
 Isomursu, Minna **II-455**  
 Isomursu, Pekka **II-455**  
 Izmalkova, Anna **II-106**
- Jaen, Javier **II-195, II-549**  
 Jain, Ajit **IV-213**  
 Jakobsen, Mikkel R. **III-455**  
 Janneck, Monique **II-274**  
 Jansen, Yvonne **III-455**  
 Järvi, Jaakko **IV-213**  
 Jensen, Iben **III-132**  
 Jeunet, Camille **I-488**  
 Johansson, Stefan **I-418**  
 Johnson, Chris **IV-663**  
 Jolaoso, Sheriff **II-463**  
 Jones, Christian Martyn **III-167**  
 Jorge, Clinton **IV-335**  
 Jorge, Joaquim **III-622**  
 Joshi, Anirudha **IV-637**  
 Jovanovic, Mladjan **III-149**  
 Jung, Ju Young **I-550**
- Kacem, Ahmed Hadj **II-115**  
 Kahl, Gerrit **IV-390**  
 Kalboussi, Anis **II-115**  
 Kappel, Karin **I-219, III-333**  
 Karousos, Nikos **I-255**  
 Katsanos, Christos **I-255**
- Katsikitis, Mary **III-167**  
 Kavšek, Branko **II-420**  
 Kawai, Hitoshi **IV-603**  
 Kejriwal, Gaurav **IV-213**  
 Kellogg, Wendy **II-38**  
 Kerne, Andruid **IV-213**  
 Keyson, David V. **IV-546**  
 Khalilbeigi, Mohammadreza **III-278**  
 Khamis, Mohamed **I-316**  
 Khan, Taimur **III-1**  
 Kharrufa, Ahmed **II-531, II-565, IV-156**  
 Kidziński, Łukasz **II-81**  
 Kim, Gerard J. **II-203, IV-10, IV-506, IV-550, IV-607**  
 Kim, Youngsun **IV-10, IV-506**  
 Kipp, Michael **IV-173**  
 Kitamura, Yoshifumi **III-296**  
 Kjeldskov, Jesper **III-410**  
 Kleemann, Mads **IV-37**  
 Kljun, Matjaž **II-420, IV-490, IV-523**  
 Knierim, Pascal **III-350**  
 Kobayashi, Marina **III-116**  
 Kojima, Akira **IV-319**  
 Kokumai, Yuji **IV-599, IV-603**  
 Köles, Máté **IV-461**  
 Komlódi, Anita **IV-461**  
 Kosmyna, Nataliya **I-506**  
 Kouyoumdjian, Alexandre **IV-148**  
 Kratky, Martina **III-333**  
 Kray, Christian **II-387**  
 Krömker, Heidi **IV-537**  
 Krüger, Antonio **III-259, IV-390**  
 Krummheuer, Antonia L. **IV-240**  
 Kubitz, Thomas **III-376**  
 Kuusinen, Kati **III-27**
- Lamas, David **II-479**  
 Lammer, Lara **II-557**  
 Lanamäki, Arto **III-9**  
 Lantz, Ann **I-418**  
 Lantz, Vuokko **IV-478, IV-611**  
 Lanzilotti, Rosa **IV-673**  
 Lárusdóttir, Marta Kristín **IV-673**  
 Latour, Thibaud **IV-79**  
 Laumer, Sven **II-38, IV-671**  
 Law, Effie Lai-Chong **I-281, III-501, IV-673**  
 Lawson, Shaun **IV-637**  
 Leat, David **II-531**  
 Lecolinet, Éric **III-221, IV-55**  
 Lee, Changhyeon **IV-607**

- Lee, Jaedong **IV-10, IV-506, IV-607**  
 Lee, Myunghee **II-203**  
 Lee, Sangwon **IV-594**  
 Lehmann, Anke **III-436**  
 Leonardi, Chiara **III-315**  
 Leonova, Anna **II-106**  
 Lepri, Bruno **III-315**  
 Li, Nan **II-81**  
 Liapis, Alexandros **I-255**  
 Lichtschlag, Leonhard **I-289**  
 Liggesmeyer, Peter **III-1**  
 Lim, Boon Pang **IV-510**  
 Lim, Mei Quin **IV-510**  
 Liu, Youming **II-611**  
 Liuska, Tiina **II-455**  
 Löffler, Diana **IV-248**  
 Lógó, Emma **IV-461**  
 Lotte, Fabien **I-472, I-488, IV-354**  
 Löw, Christian **IV-448**  
 Lubos, Paul **III-259**  
 Lucero, Andrés **II-231, III-474, IV-633**  
 Lüers, Bengt **IV-582**  
 Lutteroth, Christof **I-367**  
 Luxton-Reilly, Andrew **II-177**  
 Luyten, Kris **II-368, IV-264**  
 Luz, Nuno **I-110**  
 Lyons, Michael **IV-642**
- MacKrill, James **IV-518**  
 Madeira, Rui Neves **III-341**  
 Madsen, Sabine **III-132**  
 Mahmud, Abdullah Al **IV-542, IV-546, IV-616**  
 Makri, Stephan **IV-578**  
 Maly, Ivo **I-89**  
 Maquil, Valérie **IV-79**  
 Marinho, Amanda **IV-335**  
 Marquardt, Nicolai **II-89, IV-264, IV-644**  
 Martinie, Célia **IV-192, IV-663**  
 Martins, Bruno **IV-327**  
 Marzano, Lisa **IV-409**  
 Masclet, Cédric **I-349**  
 Masoodian, Masood **IV-657**  
 Massa, Paolo **III-315**  
 Maurer, Bernhard **II-331, IV-140**  
 Mayas, Cindy **IV-537**  
 Mayer, P. **IV-624**  
 Mazhoud, Omar **II-115**  
 Meili, Christoph **I-210**  
 Meiller, Dieter **IV-465**
- Mencarini, Eleonora **II-73**  
 Mendes, Daniel **III-622**  
 Meneweger, Thomas **III-203**  
 Merritt, Timothy **IV-37**  
 Mertl, Fabian **III-71**  
 Meschtscherjakov, Alexander **II-331, IV-621**  
 Mikovec, Zdenek **I-89**  
 Miletto, Evandro Manara **IV-590**  
 Mitchell, Alex **III-493**  
 Miura, Takahiro **III-80**  
 Moltchanova, Elena **I-263**  
 Monroe, Megan **III-418**  
 Morgan, Evan **II-47**  
 Mori, Giulio **I-165**  
 Moser, Christiane **IV-621**  
 Mossel, Annette **IV-165**  
 Mowafi, Omar **I-316**  
 Mubin, Omar **IV-542, IV-555, IV-573, IV-616**  
 Mühlhäuser, Max **III-278**  
 Muñoz, David **III-53**  
 Muñoz, Unai **I-1**  
 Murer, Martin **IV-140, IV-621, IV-659**  
 Murphy, Emma **IV-586**
- N'Kaoua, Bernard **I-488**  
 Nacher, Vicente **II-549**  
 Nagashima, Yuji **IV-502**  
 Nakano, Yukiko I. **IV-319**  
 Nefzger, Matthias **IV-300**  
 Neto, Edvar Vilar **II-395**  
 Neureiter, Katja **IV-621**  
 Nguyen, Quan **IV-173**  
 Ni, Bingxin **IV-546**  
 Nicholson, James **II-565**  
 Niculescu, Andreea I. **IV-510**  
 Niels, Adelka **II-274**  
 Nielsen, Lene **III-132**  
 Nisi, Valentina **IV-335**  
 Novikova, Jekaterina **III-239**  
 Novoa, Mauricio **IV-573**  
 Nunes, Nuno **IV-335**
- O'Hara, Kenton **III-410**  
 Obaid, Mohammad **I-263**  
 Obrist, Marianna **IV-18**  
 Ogunyemi, Abiodun **II-479**  
 Ohler, Thorsten **IV-256**  
 Ojha, Sajan Raj **III-149**

- Olaverri-Monreal, Cristina IV-626  
 Oliveira Jr., Edson IV-87  
 Olivier, Patrick II-531, II-565, IV-156  
 Olsson, Thomas III-384, IV-633  
 Omheni, Nizar II-115  
 Ostkmap, Morin II-387  
 Otsuka, Kazuhiro IV-319  
 Ozawa, Shiro IV-319
- Paier, Wolfgang IV-248  
 Palanque, Philippe II-211, IV-192, IV-637, IV-663  
 Paternò, Fabio I-165  
 Peldszus, Regina IV-663  
 Pereira, João Madeiras III-622  
 Pereira, Roberto IV-87  
 Petrie, Helen I-55, I-72, I-147, IV-665, IV-669  
 Pfeuffer, Ken II-349  
 Pherwani, Jatin IV-528  
 Pianesi, Fabio III-315  
 Piccolo, Lara S.G. I-210  
 Piedrahita-Solórzano, Giovanny III-518  
 Pillias, Clément I-531  
 Pinheiro, Mariana II-395  
 Pino, José A. III-552  
 Pirker, Michael II-412  
 Plaumann, Katrin III-526  
 Plimmer, Beryl II-177  
 Pohl, M. IV-624  
 Pohl, Norman III-376, IV-282  
 Popow, Michael IV-510  
 Porter, Joel IV-72  
 Power, Christopher I-55, I-147, III-99, IV-665  
 Prates, Raquel Oliveira I-201, IV-482, IV-667  
 Prätorius, Manuel IV-425  
 Prescher, Denise I-80  
 Preston, Anne II-531  
 Prietch, Soraia Silva I-20  
 Pucihar, Klen Čopíč II-420, IV-490, IV-523  
 Pugh, Joseph III-99  
 Qiu, Guoping II-265
- Raber, Frederic IV-390  
 Räihä, Kari-Jouko I-402  
 Rajanen, Mikko III-9  
 Rashid, Umar III-367
- Rauschenberger, Maria I-445  
 Read, Janet C. IV-655  
 Regal, Georg IV-626  
 Reijonen, Pekka II-300  
 Rello, Luz I-445  
 Ren, Gang III-239  
 Riemann, Jan III-278  
 Rioul, Olivier IV-55  
 Rist, Thomas IV-657  
 Rito, Fábio I-110  
 Rittenbruch, Markus II-134, II-150  
 Rivet, Bertrand I-506  
 Rocha, Allan I-340  
 Rogelj, Peter IV-559  
 Romano, Marco I-38  
 Romão, Teresa III-341  
 Rosson, Mary Beth I-201  
 Roudaut, Anne IV-18  
 Rovelo, Gustavo II-368  
 Roy, Quentin IV-55  
 Rudinsky, Jan II-98  
 Rukzio, Enrico III-350, III-526  
 Rutishauser, Duncan IV-573
- Sa, Ning II-249  
 Samaras, George I-523  
 Santo, Yasuhiro II-134  
 Satoh, Hironobu IV-563, IV-569  
 Sauer, Stefan IV-661  
 Savva, Andreas I-55  
 Schardong, Frederico II-89  
 Scherzinger, Aaron IV-425  
 Schiavo, Gianluca II-73  
 Schlick, Christopher IV-514  
 Schmettow, Martin IV-651  
 Schmidt, Albrecht III-376, III-402, IV-256, IV-282, IV-300  
 Schmidt, André IV-37  
 Schneegass, Stefan IV-282  
 Schnelle-Walka, Dirk IV-633  
 Schönauer, Christian IV-165  
 Schwaiger, Daniel II-412  
 Schwind, Valentin IV-282  
 Scutt, Tom IV-586  
 Seah, Sue Ann IV-18  
 Seifert, Julian III-350  
 Selker, Ted IV-37  
 Sellen, Abigail III-570  
 Seyed, Teddy II-436

- Shahid, Suleman **I-436, IV-542, IV-555, IV-573, IV-616**  
 Sharlin, Ehud **I-340, III-296**  
 Shi, Lei **IV-518**  
 Shibata, Hirohito **I-559**  
 Shibata, Kyoko **IV-533, IV-563**  
 Silveira, Milene Selbach **II-592**  
 Simonsen, Jakob Grue **III-185**  
 Sivaji, Ashok **IV-628**  
 Slegers, Karin **II-292**  
 Sliwinski, Jacek **III-167**  
 Smit, Dorothé **IV-456**  
 Snow, Stephen **II-150**  
 Soleimani, Samaneh **I-281, III-501**  
 Somanath, Sowmya **I-340**  
 Sonntag, Daniel **IV-633**  
 Sørensen, Henrik **III-410**  
 Sosik, Victoria Schwanda **III-116**  
 Sotiropoulos, Dimitris **I-255**  
 Sousa, Mario Costa **I-340**  
 Sousa, Maurício **III-622**  
 Spelmezan, Daniel **I-488**  
 Staadt, Oliver **III-436**  
 Stage, Jan **II-159**  
 Stead, Mike **IV-72**  
 Steinberger, Fabius **III-614**  
 Steinicke, Frank **III-259**  
 Stellmach, Sophie **III-570**  
 Stelzer, Anselmo **IV-537**  
 Stock, Oliviero **II-73**  
 Strohmeier, Paul **I-332**  
 Sturm, Christian **IV-630**  
 Subramanian, Sriram **I-488, IV-18**  
 Sutherland, Alistair **IV-499**  
 Sutherland, Craig J. **II-177**  
 Sutton, Selina **IV-156**  
 Swallow, David **I-147**  
 Szabó, Bálint **IV-461**
- Takano, Kentaro **I-559**  
 Takashima, Kazuki **III-296**  
 Takashina, Tomomi **IV-599, IV-603**  
 Tang, Anthony **II-89, II-436**  
 Tano, Shun'ichi **I-559**  
 Tarkkanen, Kimmo **II-300**  
 Tarpin-Bernard, Franck **I-506**  
 Tausch, Sarah **III-614**  
 Teichrieb, Veronica **II-395**  
 Tellioglu, Hilda **IV-448, IV-624**  
 Terauchi, Mina **IV-502**
- Thies, William **II-505**  
 Thomaschewski, Jörg **I-445**  
 Timmermann, Janko **IV-582**  
 Tobias, Eric **IV-79**  
 Toriiyuka, Takashi **IV-248**  
 Trösterer, Sandra **II-331**  
 Tscheligi, Manfred **II-331, III-203, IV-140, IV-621, IV-626, IV-659**  
 Turner, Jayson **II-315**  
 Turunen, Markku **IV-633**
- Vääänänen-Vainio-Mattila, Kaisa **III-384**  
 Vaish, Rajan **II-505**  
 Valencia, Xabier **I-1**  
 van der Veer, Gerrit **IV-675**  
 Vanacken, Davy **II-368**  
 Vandenberghe, Bert **IV-417**  
 Vashistha, Aditya **II-505**  
 Vasiliou, Christina **II-55**  
 Vatavu, Radu-Daniel **IV-1, IV-165**  
 Väyrynen, Jani **IV-363**  
 Velloso, Eduardo **I-384, II-315**  
 Vermeulen, Jo **IV-264**  
 Vertegaal, Roel **I-332**  
 Vi, Chi **I-488**  
 Vičić, Jernej **II-420**  
 Vincze, Markus **II-557**  
 von Zezschwitz, Emanuel **II-428**  
 Vormann, Anja **III-71**
- Wacker, Philipp **I-289**  
 Wagner, Johannes **I-263**  
 Wagner, Julie **IV-231**  
 Wainohu, Derek **IV-573**  
 Walbaum, Katrin **IV-514**  
 Walldius, Åke **I-298**  
 Wang, Jingtao **II-611**  
 Watkins, Matt **IV-586**  
 Watts, Leon **III-239**  
 Weber, Gerald **I-367**  
 Weber, Gerhard **I-80, IV-469, IV-665**  
 Webster, Gemma **IV-677**  
 Weiss, Astrid **II-557**  
 Whittle, Jon **III-367**  
 Wibowo, Seno A. **IV-510**  
 Wilhelm, Dennis **II-387**  
 Wilkins, Jason **IV-213**  
 Wimmer, Christoph **I-219, III-333**  
 Winckler, Marco **II-211, IV-661**  
 Winkler, Christian **III-350**

- Winschiers-Theophilus, Heike [II-489](#)  
Wippoo, Meia [I-210](#)  
Wittland, Jan [I-453](#)  
Wobrock, Dennis [I-472](#)  
Wolf, Katrin [III-376, IV-282](#)  
Wortelen, Bertram [IV-105](#)  
Wuchse, Martin [II-331](#)  
Wurhofer, Daniela [III-203](#)  
Wybrands, Marius [IV-582](#)  
  
Xenos, Michalis [I-255](#)  
  
Yabu, Ken-ichiro [III-80](#)  
Yamaguchi, Takumi [IV-569](#)  
Yamashita, Naomi [I-183](#)  
Yang, Huahai [II-249](#)  
Yang, Xing-Dong [II-436](#)  
Yeo, Kheng Hui [IV-510](#)  
Yeoh, Ken Neth [I-367](#)  
Yokoyama, Hitomi [III-296](#)  
  
Yoshida, Masanobu [IV-569](#)  
Yu, Bin [III-45](#)  
Yu, Jaeun [IV-550](#)  
Yuan, Xiaojun [II-249](#)  
Yuras, Gabriel [III-526](#)  
  
Zagler, W. [IV-624](#)  
Zaīti, Ionuț-Alexandru [IV-165](#)  
Zaman, Tariq [II-489](#)  
Zancanaro, Massimo [II-73, III-315](#)  
Zaphiris, Panayiotis [II-55](#)  
Zeng, Limin [IV-469](#)  
Zhang, Min [II-265](#)  
Zhang, Yanxia [III-570](#)  
Zhou, Jianlong [I-550](#)  
Zhu, Bin [II-436](#)  
Ziefle, Martina [I-289, I-453](#)  
Ziegler, Jürgen [II-20](#)  
Zlokazova, Tatiana [II-106](#)