

Julie A. Jacko (Ed.)

LNCS 6762

Human-Computer Interaction

Interaction Techniques and Environments

14th International Conference, HCI International 2011
Orlando, FL, USA, July 2011
Proceedings, Part II

2
Part II



Springer

Commenced Publication in 1973

Founding and Former Series Editors:

Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

David Hutchison

Lancaster University, UK

Takeo Kanade

Carnegie Mellon University, Pittsburgh, PA, USA

Josef Kittler

University of Surrey, Guildford, UK

Jon M. Kleinberg

Cornell University, Ithaca, NY, USA

Alfred Kobsa

University of California, Irvine, CA, USA

Friedemann Mattern

ETH Zurich, Switzerland

John C. Mitchell

Stanford University, CA, USA

Moni Naor

Weizmann Institute of Science, Rehovot, Israel

Oscar Nierstrasz

University of Bern, Switzerland

C. Pandu Rangan

Indian Institute of Technology, Madras, India

Bernhard Steffen

TU Dortmund University, Germany

Madhu Sudan

Microsoft Research, Cambridge, MA, USA

Demetri Terzopoulos

University of California, Los Angeles, CA, USA

Doug Tygar

University of California, Berkeley, CA, USA

Gerhard Weikum

Max Planck Institute for Informatics, Saarbruecken, Germany

Julie A. Jacko (Ed.)

Human-Computer Interaction

Interaction Techniques and Environments

14th International Conference, HCI International 2011
Orlando, FL, USA, July 9-14, 2011
Proceedings, Part II



Springer

Volume Editor

Julie A. Jacko
University of Minnesota
School of Public Health and Institute for Health Informatics
1260 Mayo (MMC 807), 420 Delaware Street S.E.
Minneapolis, MN 55455, USA
E-mail: jacko@umn.edu

ISSN 0302-9743
ISBN 978-3-642-21604-6
DOI 10.1007/978-3-642-21605-3
Springer Heidelberg Dordrecht London New York

e-ISSN 1611-3349
e-ISBN 978-3-642-21605-3

Library of Congress Control Number: 2011929076

CR Subject Classification (1998): H.5.2, H.5, H.4, I.2.10, I.4-5, C.2

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

© Springer-Verlag Berlin Heidelberg 2011
This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

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

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

Printed on acid-free paper

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

Foreword

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

A total of 4,039 individuals from academia, research institutes, industry and governmental agencies from 67 countries submitted contributions, and 1,318 papers that were judged to be of high scientific quality were included in the program. These papers address the latest research and development efforts and highlight the human aspects of design and use of computing systems. The papers accepted for presentation thoroughly cover the entire field of human–computer interaction, addressing major advances in knowledge and effective use of computers in a variety of application areas.

This volume, edited by Julie A. Jacko, contains papers in the thematic area of human–computer interaction (HCI), addressing the following major topics:

- Touch-based and haptic interaction
- Gaze and gesture-based interaction
- Voice, natural language and dialogue
- Novel interaction techniques and devices
- Avatars and embodied interaction

The remaining volumes of the HCI International 2011 Proceedings are:

- Volume 1, LNCS 6761, Human–Computer Interaction—Design and Development Approaches (Part I), edited by Julie A. Jacko
- Volume 3, LNCS 6763, Human–Computer Interaction—Towards Mobile and Intelligent Interaction Environments (Part III), edited by Julie A. Jacko
- Volume 4, LNCS 6764, Human–Computer Interaction—Users and Applications (Part IV), edited by Julie A. Jacko
- Volume 5, LNCS 6765, Universal Access in Human–Computer Interaction—Design for All and eInclusion (Part I), edited by Constantine Stephanidis
- Volume 6, LNCS 6766, Universal Access in Human–Computer Interaction—Users Diversity (Part II), edited by Constantine Stephanidis
- Volume 7, LNCS 6767, Universal Access in Human–Computer Interaction—Context Diversity (Part III), edited by Constantine Stephanidis

- Volume 8, LNCS 6768, Universal Access in Human–Computer Interaction—Applications and Services (Part IV), edited by Constantine Stephanidis
- Volume 9, LNCS 6769, Design, User Experience, and Usability—Theory, Methods, Tools and Practice (Part I), edited by Aaron Marcus
- Volume 10, LNCS 6770, Design, User Experience, and Usability—Understanding the User Experience (Part II), edited by Aaron Marcus
- Volume 11, LNCS 6771, Human Interface and the Management of Information—Design and Interaction (Part I), edited by Michael J. Smith and Gavriel Salvendy
- Volume 12, LNCS 6772, Human Interface and the Management of Information—Interacting with Information (Part II), edited by Gavriel Salvendy and Michael J. Smith
- Volume 13, LNCS 6773, Virtual and Mixed Reality—New Trends (Part I), edited by Randall Shumaker
- Volume 14, LNCS 6774, Virtual and Mixed Reality—Systems and Applications (Part II), edited by Randall Shumaker
- Volume 15, LNCS 6775, Internationalization, Design and Global Development, edited by P.L. Patrick Rau
- Volume 16, LNCS 6776, Human-Centered Design, edited by Masaaki Kurosu
- Volume 17, LNCS 6777, Digital Human Modeling, edited by Vincent G. Duffy
- Volume 18, LNCS 6778, Online Communities and Social Computing, edited by A. Ant Ozok and Panayiotis Zaphiris
- Volume 19, LNCS 6779, Ergonomics and Health Aspects of Work with Computers, edited by Michelle M. Robertson
- Volume 20, LNAI 6780, Foundations of Augmented Cognition: Directing the Future of Adaptive Systems, edited by Dylan D. Schmorrow and Cali M. Fidopiastis
- Volume 21, LNAI 6781, Engineering Psychology and Cognitive Ergonomics, edited by Don Harris
- Volume 22, CCIS 173, HCI International 2011 Posters Proceedings (Part I), edited by Constantine Stephanidis
- Volume 23, CCIS 174, HCI International 2011 Posters Proceedings (Part II), edited by Constantine Stephanidis

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

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

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

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

July 2011

Constantine Stephanidis

Organization

Ergonomics and Health Aspects of Work with Computers

Program Chair: Michelle M. Robertson

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

Human Interface and the Management of Information

Program Chair: Michael J. Smith

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

Human–Computer Interaction

Program Chair: Julie A. Jacko

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

Engineering Psychology and Cognitive Ergonomics

Program Chair: Don Harris

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

Universal Access in Human–Computer Interaction

Program Chair: Constantine Stephanidis

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

Zhengjie Liu, P.R. China
Klaus Miesenberger, Austria
Helen Petrie, UK
Michael Pieper, Germany
Anthony Savidis, Greece
Andrew Sears, USA
Christian Stary, Austria

Hirotada Ueda, Japan
Jean Vanderdonckt, Belgium
Gregg C. Vanderheiden, USA
Gerhard Weber, Germany
Harald Weber, Germany
Panayiotis Zaphiris, Cyprus

Virtual and Mixed Reality

Program Chair: Randall Shumaker

Pat Banerjee, USA
Mark Billinghurst, New Zealand
Charles E. Hughes, USA
Simon Julier, UK
David Kaber, USA
Hirokazu Kato, Japan
Robert S. Kennedy, USA
Young J. Kim, Korea
Ben Lawson, USA
Gordon McK Mair, UK

David Pratt, UK
Albert "Skip" Rizzo, USA
Lawrence Rosenblum, USA
Jose San Martin, Spain
Dieter Schmalstieg, Austria
Dylan Schmorrow, USA
Kay Stanney, USA
Janet Weisenford, USA
Mark Wiederhold, USA

Internationalization, Design and Global Development

Program Chair: P.L. Patrick Rau

Michael L. Best, USA
Alan Chan, Hong Kong
Lin-Lin Chen, Taiwan
Andy M. Dearden, UK
Susan M. Dray, USA
Henry Been-Lirn Duh, Singapore
Vanessa Evers, The Netherlands
Paul Fu, USA
Emilie Gould, USA
Sung H. Han, Korea
Veikko Ikonen, Finland
Toshikazu Kato, Japan
Esin Kiris, USA
Apala Lahiri Chavan, India

James R. Lewis, USA
James J.W. Lin, USA
Rungtai Lin, Taiwan
Zhengjie Liu, P.R. China
Aaron Marcus, USA
Allen E. Milewski, USA
Katsuhiko Ogawa, Japan
Oguzhan Ozcan, Turkey
Girish Prabhu, India
Kerstin Röse, Germany
Supriya Singh, Australia
Alvin W. Yeo, Malaysia
Hsiu-Ping Yueh, Taiwan

Online Communities and Social Computing

Program Chairs: A. Ant Ozok, Panayiotis Zaphiris

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

Augmented Cognition

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

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

Digital Human Modeling

Program Chair: Vincent G. Duffy

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

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

Human-Centered Design

Program Chair: Masaaki Kurosu

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

Design, User Experience, and Usability

Program Chair: Aaron Marcus

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

XIV Organization

Rüdiger Heimgärtner, Germany
Brigitte Herrmann, Germany
Sabine Kabel-Eckes, USA
Kaleem Khan, Canada
Jonathan Kies, USA
Jon Kolko, USA
Helga Letowt-Vorbek, South Africa
James Lin, USA
Frazer McKimm, Ireland
Michael Renner, Switzerland

Christine Ronnewinkel, Germany
Elizabeth Rosenzweig, USA
Paul Sherman, USA
Ben Shneiderman, USA
Christian Sturm, Germany
Brian Sullivan, USA
Jaakko Villa, Finland
Michele Visciola, Italy
Susan Weinschenk, USA

HCI International 2013

The 15th International Conference on Human–Computer Interaction, HCI International 2013, will be held jointly with the affiliated conferences in the summer of 2013. It will cover a broad spectrum of themes related to human–computer interaction (HCI), including theoretical issues, methods, tools, processes and case studies in HCI design, as well as novel interaction techniques, interfaces and applications. The proceedings will be published by Springer. More information about the topics, as well as the venue and dates of the conference, will be announced through the HCI International Conference series website: <http://www.hci-international.org/>

General Chair

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

Table of Contents – Part II

Part I: Touch-Based and Haptic Interaction

Development of a High Definition Haptic Rendering for Stability and Fidelity	3
<i>Katsuhito Akahane, Takeo Hamada, Takehiko Yamaguchi, and Makoto Sato</i>	
Designing a Better Morning: A Study on Large Scale Touch Interface Design	13
<i>Onur Asan, Mark Omernick, Dain Peer, and Enid Montague</i>	
Experimental Evaluations of Touch Interaction Considering Automotive Requirements	23
<i>Andreas Haslbeck, Severina Popova, Michael Krause, Katrina Pecot, Jürgen Mayer, and Klaus Bengler</i>	
More Than Speed? An Empirical Study of Touchscreens and Body Awareness on an Object Manipulation Task	33
<i>Rachelle Kristof Hippler, Dale S. Klopfer, Laura Marie Leventhal, G. Michael Poor, Brandi A. Klein, and Samuel D. Jaffee</i>	
TiMBA – Tangible User Interface for Model Building and Analysis	43
<i>Chih-Pin Hsiao and Brian R. Johnson</i>	
Musical Skin: A Dynamic Interface for Musical Performance	53
<i>Heng Jiang, Teng-Wen Chang, and Cha-Lin Liu</i>	
Analyzing User Behavior within a Haptic System	62
<i>Steve Johnson, Yueqing Li, Chang Soo Nam, and Takehiko Yamaguchi</i>	
Usability Testing of the Interaction of Novices with a Multi-touch Table in Semi Public Space	71
<i>Markus Jokisch, Thomas Bartoschek, and Angela Schwering</i>	
Niboshi for Slate Devices: A Japanese Input Method Using Multi-touch for Slate Devices	81
<i>Gimpei Kimioka, Buntarou Shizuki, and Jiro Tanaka</i>	
An Investigation on Requirements for Co-located Group-Work Using Multitouch-, Pen-Based- and Tangible-Interaction	90
<i>Karsten Nebe, Tobias Müller, and Florian Klompmaker</i>	

XVIII Table of Contents – Part II

Exploiting New Interaction Techniques for Disaster Control Management Using Multitouch-, Tangible- and Pen-Based-Interaction	100
<i>Karsten Nebe, Florian Klompmaker, Helge Jung, and Holger Fischer</i>	
Saving and Restoring Mechanisms for Tangible User Interfaces through Tangible Active Objects	110
<i>Eckard Riedenklau, Thomas Hermann, and Helge Ritter</i>	
Needle Insertion Simulator with Haptic Feedback	119
<i>Seungjae Shin, Wanjoo Park, Hyunchul Cho, Sehyung Park, and Laehyun Kim</i>	
Measurement of Driver's Distraction for an Early Prove of Concepts in Automotive Industry at the Example of the Development of a Haptic Touchpad	125
<i>Roland Spies, Andreas Blattner, Christian Lange, Martin Wohlfarter, Klaus Bengler, and Werner Hamberger</i>	
A Tabletop-Based Real-World-Oriented Interface.....	133
<i>Hiroshi Takeda, Hidetoshi Miyao, Minoru Maruyama, and David Asano</i>	
What You Feel Is What I Do: A Study of Dynamic Haptic Interaction in Distributed Collaborative Virtual Environment	140
<i>Sehat Ullah, Xianging Liu, Samir Otmane, Paul Richard, and Malik Mallem</i>	
A Framework Interweaving Tangible Objects, Surfaces and Spaces	148
<i>Andy Wu, Jayraj Jog, Sam Mendenhall, and Ali Mazalek</i>	
The Effect of Haptic Cues on Working Memory in 3D Menu Selection	158
<i>Takehiko Yamaguchi, Damien Chamaret, and Paul Richard</i>	

Part II: Gaze and Gesture-Based Interaction

Face Recognition Using Local Graph Structure (LGS)	169
<i>Eimad E.A. Abusham and Housam K. Bashir</i>	
Eye-gaze Detection by Image Analysis under Natural Light	176
<i>Kiyohiko Abe, Shoichi Ohi, and Minoru Ohyama</i>	
Multi-user Pointing and Gesture Interaction for Large Screen Using Infrared Emitters and Accelerometers	185
<i>Leonardo Angelini, Maurizio Caon, Stefano Carrino, Omar Abou Khaled, and Elena Mugellini</i>	

Gesture Identification Based on Zone Entry and Axis Crossing.....	194
<i>Ryosuke Aoki, Yutaka Karatsu, Masayuki Ihara, Atsuhiko Maeda, Minoru Kobayashi, and Shingo Kagami</i>	
Attentive User Interface for Interaction within Virtual Reality Environments Based on Gaze Analysis.....	204
<i>Florin Barbuceanu, Csaba Antonya, Mihai Duguleana, and Zoltan Rusak</i>	
A Low-Cost Natural User Interaction Based on a Camera Hand-Gestures Recognizer	214
<i>Mohamed-Ikbel Boulaibar, Thomas Burger, Franck Poirier, and Gilles Coppin</i>	
Head-Computer Interface: A Multimodal Approach to Navigate through Real and Virtual Worlds	222
<i>Francesco Carrino, Julien Tscherrig, Elena Mugellini, Omar Abou Khaled, and Rolf Ingold</i>	
3D-Position Estimation for Hand Gesture Interface Using a Single Camera	231
<i>Seung-Hwan Choi, Ji-Hyeong Han, and Jong-Hwan Kim</i>	
Hand Gesture for Taking Self Portrait	238
<i>Shaowei Chu and Jiro Tanaka</i>	
Hidden-Markov-Model-Based Hand Gesture Recognition Techniques Used for a Human-Robot Interaction System	248
<i>Chin-Shyurng Fahn and Keng-Yu Chu</i>	
Manual and Accelerometer Analysis of Head Nodding Patterns in Goal-oriented Dialogues.....	259
<i>Masashi Inoue, Toshio Irino, Nobuhiro Furuyama, Ryoko Hanada, Takako Ichinomiya, and Hiroyasu Massaki</i>	
Facial Expression Recognition Using AAMICPF	268
<i>Jun-Sung Lee, Chi-Min Oh, and Chil-Woo Lee</i>	
Verification of Two Models of Ballistic Movements	275
<i>Jui-Feng Lin and Colin G. Drury</i>	
Gesture Based Automating Household Appliances	285
<i>Wei Lun Ng, Chee Kyun Ng, Nor Kamariah Noordin, and Borhanuddin Mohd. Ali</i>	
Upper Body Gesture Recognition for Human-Robot Interaction	294
<i>Chi-Min Oh, Md. Zahidul Islam, Jun-Sung Lee, Chil-Woo Lee, and In-So Kweon</i>	

Gaze-Directed Hands-Free Interface for Mobile Interaction	304
<i>Gie-seo Park, Jong-gil Ahn, and Gerard J. Kim</i>	
Eye-Movement-Based Instantaneous Cognition Model for Non-verbal Smooth Closed Figures	314
<i>Yuzo Takahashi and Shoko Koshi</i>	

Part III: Voice, Natural Language and Dialogue

VOSS -A Voice Operated Suite for the Barbadian Vernacular	325
<i>David Byer and Colin Depradine</i>	
New Techniques for Merging Text Versions	331
<i>Darius Dadgari and Wolfgang Stuerzlinger</i>	
Modeling the Rhetoric of Human-Computer Interaction	341
<i>Iris Howley and Carolyn Penstein Rosé</i>	
Recommendation System Based on Interaction with Multiple Agents for Users with Vague Intention	351
<i>Itaru Kuramoto, Atsushi Yasuda, Mitsuru Minakuchi, and Yoshihiro Tsujino</i>	
A Review of Personality in Voice-Based Man Machine Interaction	358
<i>Florian Metze, Alan Black, and Tim Polzehl</i>	
Can Indicating Translation Accuracy Encourage People to Rectify Inaccurate Translations?	368
<i>Mai Miyabe and Takashi Yoshino</i>	
Design of a Face-to-Face Multilingual Communication System for a Handheld Device in the Medical Field	378
<i>Shun Ozaki, Takuo Matsunobe, Takashi Yoshino, and Aguri Shigeno</i>	
Computer Assistance in Bilingual Task-Oriented Human-Human Dialogues	387
<i>Sven Schmeier, Matthias Rebel, and Renlong Ai</i>	
Developing and Exploiting a Multilingual Grammar for Human-Computer Interaction	396
<i>Xian Zhang, Rico Andrich, and Dietmar Rösner</i>	

Part IV: Novel Interaction Techniques and Devices

Dancing Skin: An Interactive Device for Motion	409
<i>Sheng-Han Chen, Teng-Wen Chang, and Sheng-Cheng Shih</i>	
A Hybrid Brain-Computer Interface for Smart Home Control	417
<i>Günter Edlinger, Clemens Holzner, and Christoph Guger</i>	

Integrated Context-Aware and Cloud-Based Adaptive Home Screens for Android Phones	427
<i>Tor-Morten Grønli, Jarle Hansen, and Gheorghita Ghinea</i>	
Evaluation of User Support of a Hemispherical Sub-Display with GUI Pointing Functions	436
<i>Shinichi Ike, Saya Yokoyama, Yuya Yamanishi, Naohisa Matsuuchi, Kazunori Shimamura, Takumi Yamaguchi, and Haruya Shiba</i>	
Uni-model Human System Interface Using sEMG	446
<i>Srinivasan Jayaraman and Venkatesh Balasubramanian</i>	
An Assistive Bi-modal User Interface Integrating Multi-channel Speech Recognition and Computer Vision	454
<i>Alexey Karpov, Andrey Ronzhin, and Irina Kipyatkova</i>	
A Method of Multiple Odors Detection and Recognition	464
<i>Dong-Kyu Kim, Yong-Wan Roh, and Kwang-Seok Hong</i>	
Report on a Preliminary Study Using Breath Control and a Virtual Jogging Scenario as Biofeedback for Resilience Training	474
<i>Jacquelyn Ford Morie, Eric Chance, and J. Galen Buckwalter</i>	
Low Power Wireless EEG Headset for BCI Applications	481
<i>Shrishail Patki, Bernard Grundlehner, Toru Nakada, and Julien Penders</i>	
Virtual Mouse: A Low Cost Proximity-Based Gestural Pointing Device	491
<i>Sheng Kai Tang, Wen Chieh Tseng, Wei Wen Luo, Kuo Chung Chiu, Sheng Ta Lin, and Yen Ping Liu</i>	
Innovative User Interfaces for Wearable Computers in Real Augmented Environment	500
<i>Yun Zhou, Bertrand David, and René Chalon</i>	
Part V: Avatars and Embodied Interaction	
Influence of Prior Knowledge and Embodiment on Human-Agent Interaction	513
<i>Yugo Hayashi, Victor V. Kryssanov, Kazuhisa Miwa, and Hitoshi Ogawa</i>	
The Effect of Physical Embodiment of an Animal Robot on Affective Prosody Recognition	523
<i>Myounghoon Jeon and Infantdani A. Rayan</i>	

Older User-Computer Interaction on the Internet: How Conversational Agents Can Help	533
<i>Wi-Suk Kwon, Veena Chattaraman, Soo In Shim, Hanen Alnizami, and Juan Gilbert</i>	
An Avatar-Based Help System for Web-Portals	537
<i>Helmut Lang, Christian Mosch, Bastian Boegel, David Michel Benoit, and Wolfgang Minker</i>	
mediRobbi: An Interactive Companion for Pediatric Patients during Hospital Visit	547
<i>Szu-Chia Lu, Nicole Blackwell, and Ellen Yi-Luen Do</i>	
Design of Shadows on the OHP Metaphor-Based Presentation Interface Which Visualizes a Presenter's Actions	557
<i>Yuichi Murata, Kazutaka Kurihara, Toshio Mochizuki, Buntarou Shizuki, and Jiro Tanaka</i>	
Web-Based Nonverbal Communication Interface Using 3DAgents with Natural Gestures	565
<i>Toshiya Naka and Toru Ishida</i>	
Taking Turns in Flying with a Virtual Wingman	575
<i>Pim Nauts, Willem van Doesburg, Emiel Krahmer, and Anita Cremers</i>	
A Configuration Method of Visual Media by Using Characters of Audiences for Embodied Sport Cheering	585
<i>Kentaro Okamoto, Michiya Yamamoto, and Tomio Watanabe</i>	
Introducing Animatronics to HCI: Extending Reality-Based Interaction	593
<i>G. Michael Poor and Robert J.K. Jacob</i>	
Development of Embodied Visual Effects Which Expand the Presentation Motion of Emphasis and Indication	603
<i>Yuya Takao, Michiya Yamamoto, and Tomio Watanabe</i>	
Experimental Study on Appropriate Reality of Agents as a Multi-modal Interface for Human-Computer Interaction	613
<i>Kaori Tanaka, Tatsunori Matsui, and Kazuaki Kojima</i>	
Author Index	623

Part I

Touch-Based and Haptic Interaction

Development of a High Definition Haptic Rendering for Stability and Fidelity

Katsuhito Akahane¹, Takeo Hamada¹, Takehiko Yamaguchi², and Makoto Sato¹

¹ Precision & Intelligence Lab

Tokyo Institute of Technology, Japan

² Laboratoire d'Ingénierie des Systèmes Automatisés (LISA),
Université d'Angers, Angers, France

kakahane@hi.pi.titech.ac.jp, hamada.t.ad@m.titech.ac.jp,
takehiko.yamaguchi@univ-angers.fr, msato@pi.titech.ac.jp

Abstract. In this study, we developed and evaluated a 10kHz high definition haptic rendering system which could display at real-time video-rate (60Hz) for general VR applications. Our proposal required both fidelity and stability in a multi-rate system, with a frequency ratio of approximately 160 times. To satisfy these two criteria, there were some problems to be resolved. To achieve only stability, we could use a virtual coupling method to link a haptic display and a virtual object. However, due to its low coupling impedance, this method is not good for realization of fidelity and quality of manipulation. Therefore, we developed a multi-rate system with two level up-samplings for both fidelity and stability of haptic sensation. The first level up-sampling achieved stability by the virtual coupling, and the second level achieved fidelity by 10kHz haptic rendering to compensate for the haptic quality lost from the coupling process. We confirmed that, with our proposed system, we could achieve both stability and fidelity of haptic rendering through a computer simulation and a 6DOF haptic interface (SPIDAR-G) with a rigid object simulation engine.

Keywords: Haptic interface, High definition haptic and SPIDAR.

1 Introduction

Over the past few decades, a considerable number of studies have been conducted on haptic interfaces. Interest in VR (Virtual Reality) applications that have physics simulators with haptic interfaces have, in particular, been growing. Generally, it is thought that a haptic interfaces should be controlled at approximately 1kHz frequency to achieve stable haptic sensation. However, there seems not to be enough haptic quality in the 1kHz haptic rendering to display hard surfaces with high fidelity. It is also difficult to accurately maintain the 1kHz frequency on a general PC (personal computer) environment where VR applications need a large amount of resources, depending on the scale of the application. We developed a special haptic controller that achieved a 10kHz high definition haptic rendering in a 3DOF (Degree Of Freedom) haptic interface [11].

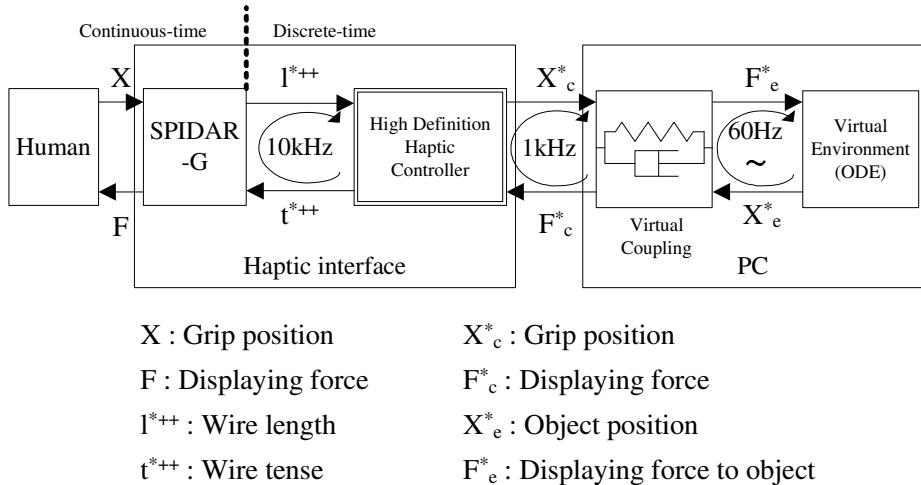


Fig. 1. System configuration

The results of the study indicated that the haptic ability (Z-width) was approximately ten times as large as the ability of a traditional 1kHz haptic rendering. It was possible to stably display a hard surface with high fidelity in a VR application with a penalty based method. While we'd like to use an analytical based method [3] [16] [17] which allows a video-rate control frequency (60Hz) for visual sensation in a VR application, the frequency difference between the systems (60Hz to 10kHz) is approximately 160 times. One of the most important issues in this study was how to interpolate the large difference in satisfying both stability and fidelity. We describe these issues in following sections.

2 Achieving Stability and Fidelity

We aimed for achieving both stability and fidelity in manipulating a virtual rigid-object. The purposes of this study are given below.

The purposes of this study at visual sensation

- To avoid penetration between rigid-objects (fidelity of visual sensation).
- To maintain stability against a haptic device.
- To use a VR application frequency over video-rate(60Hz–).

The purposes of this study at haptic sensation

- To display the extra-stiffness of a virtual rigid-object.
- To maintain stability against a video-rate VR application.
- To use a haptic rendering frequency over 10kHz.

It is important to define an objective criterion for fidelity. Contact feeling is one of the most important factors in manipulation of a virtual rigid-object. When virtual rigid-objects collided with each other, a haptic device needs to display a life-like

force. In this study, fidelity of manipulation of a virtual rigid object was evaluated by the stiffness of the object surface at the contact point. When we chose a penalty based method [19] [18] for a VR application in order to use an impedance haptic display, we could not avoid penetration between colliding objects when calculating the interaction force. From the viewpoint of visual sensation, such penetration influence is unreasonable because penetration between rigid-objects does not occur in the real world. On the other hand, with an analytical based method, such penetration does not occur and a video-rate frequency while conducting the interaction of objects can be stably maintained. Therefore, an analytical based method for VR application was the proper choice to achieve the purposes in this study. We adopted an analytical based method despite it having some problems with usage in an impedance haptic display.

3 Haptic Display

In this study, we focused on an impedance haptic display. There are several studies about controlling a haptic display for stability. Generally, it is thought that a high control frequency is necessary to achieve stable haptic sensation [15] [6] [14]. Colgate et al. [6] obtained a relationship (1) among the virtual wall's impedance (K and B), a haptic device inherent viscosity (b) and a control sampling time (T) to display haptic sensation for stability with a passivity method.

$$b > \frac{KT}{2} + B \quad (1)$$

According to the relationship (1), when trying to display a high impedance virtual wall, we need to decrease sampling time (T) (increase sampling frequency) or increase the inherent device dumper (b). When we increase the inherent device viscosity, it decreases the transparency between the device and user. Therefore, to increase sampling frequency, it is important to achieve stability and fidelity at the device level.

4 Virtual Coupling

When we use an impedance display in an analytical VR application, we cannot connect the two systems directly. Colgate et al. [7] proposed a coupling between a haptic device and a VR application with a virtual spring and dumper. The structure linking the two systems is called a virtual coupling [7] [8] [20]. Fig. 2 shows the connection of an impedance haptic display and an analytical VR application with virtual coupling. In this study, this coupling process allowed up sampling from video-rate (approximately 60Hz) to transmission frequency (approximately 1kHz) between a PC and a haptic controller. The force (F_e) inputted into the VR application was adopted to the average along time (\bar{F}_h) of the force (F_h) inputted into the haptic device.

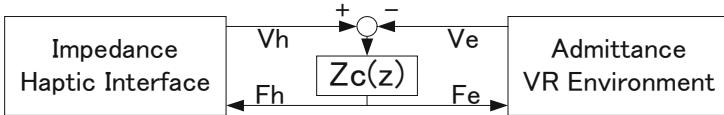


Fig. 2. Virtual coupling

$$F_h = Z_c(V_h - V_e) \quad (2)$$

$$F_e = \overline{F}_h \quad (3)$$

In this coupling structure, the stability and fidelity of the coupling depends on the coupling impedance (Z_c). To increase the fidelity, coupling impedance must also be increased. However, with VR applications running at video-rate, it is impossible to increase the impedance to maintain stability because of the loss of passivity of the virtual impedance causing severe oscillations in the haptic end-effector for the user. In actuality, the coupling impedance must be set to a low value when using this coupling, though fidelity of manipulation of the virtual object suffers. In this study, to compensate for the loss of quality in the coupling process, a high definition haptic rendering with interpolation from information of a displaying wall was adopted.

5 High Definition Haptic Rendering

Most research about multi rate simulations in haptic sensation are directed for 1kHz frequency [1] [9] [10] [5] [4]. To achieve 10kHz frequency, an interpolating algorithm needs to be as simple as possible [11]. Especially in a 6DOF haptic rendering, we need the shape and dynamics of a virtual object which are connected to a haptic end-effector to calculate displaying force accurately. In the case of 10kHz rendering, the difference of the position of the haptic end-effector is small. We can make the accuracy of interpolation simpler with the basis of two assumptions during an interpolation. First, the dynamics of a virtual object linked to a haptic end-effector are static. Second, the rendering surface of the virtual object is constant.

5.1 6DOF 10kHz Haptic Interpolation

We proposed a 6DOF interpolating algorithm for the 10kHz high definition haptic rendering. This interpolating haptic rendering (IHR) did not need a shape of a virtual object to calculate displaying force. The interpolating force was calculated by the force of the rendering surface and an impedance of the rendering surface. Fig. 3 indicates the interpolating haptic rendering in a 3DOF translation force. For a 3DOF rotation force, an analogy of the 3DOF translation was adopted.

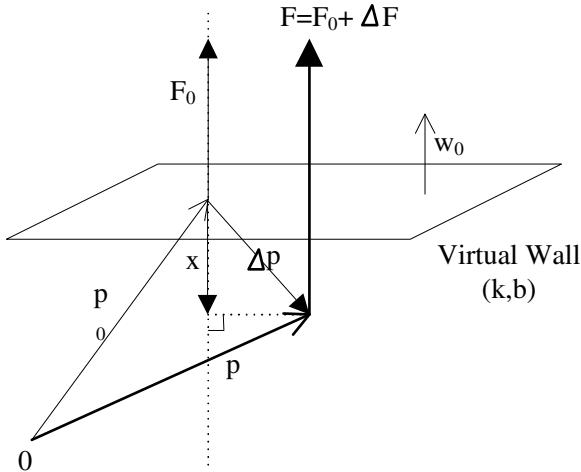


Fig. 3. Interpolating haptic rendering (IHR)

5.2 High Definition Haptic Controller

As described in previous sections, a high definition haptic controller (HDHC) was developed [11]. The controller has an SH4 micro-processor (SH7750@200MHz) made by Renesas Technology and a USB2.0 (universal serial bus) interface. The processor conducts haptic device inherent calculations such as measuring a position of the haptic end-effector from the encoders count and distributing a force to the actuators. In order to achieve a high control frequency (10kHz), it was important to reduce overhead time in calling interrupt functions. Instead of using an OS (operating system) or RTOS (real time operating system) on the processor, the source code programming was implemented as the native programming. In particular, the time critical functions were written in assembler code. For the development environment, the implementation was used with HEW3 (Renesas Technology) which consists of a C/C++ compiler assembler (SuperH RISC engine tool-chain 7.1.3.0).

6 System Configuration

To summarize the system configuration (Fig. 4), a VR application based on an analytical method was carried out at video-rate (60Hz). A 6DOF haptic device based on an impedance display was performed at 10kHz with HDHC. The whole system was conducted as a multi rate system with two level up-samplings with a frequency ratio of approximately 160 times (60Hz over 10kHz). The first up-sampling process with a frequency ratio of 60Hz (video-rate) over 1kHz (USB transmission speed) consisted of the virtual coupling which connects the different input-output systems between the VR application and the haptic device. This process allowed for stable manipulation of a virtual object with a low coupling impedance. The second up-sampling process with a frequency ratio of 1kHz over 10kHz consisted of the high

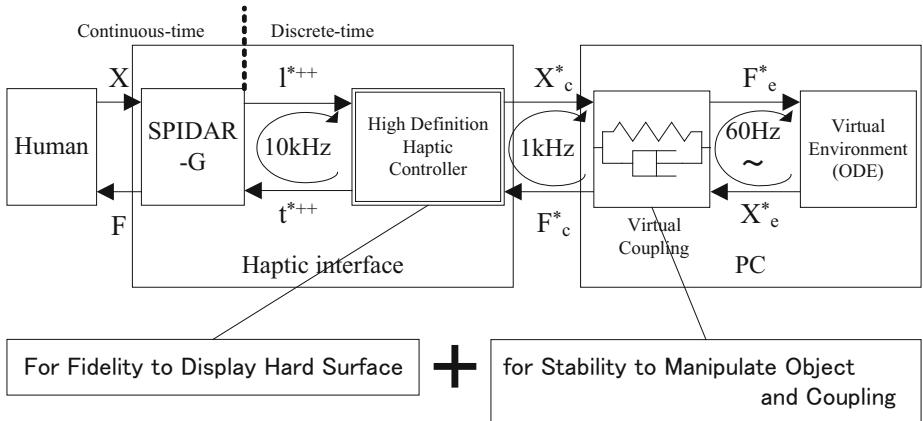


Fig. 4. System configuration

definition haptic rendering with the interpolation from information of the displaying wall. This process allowed for manipulation of a virtual object with high fidelity.

7 Evaluation

In this section, we describe an evaluation of our proposed system. We examined it with a computer simulation and an application with a rigid object simulation engine.

7.1 Computer Simulation

We performed a computer simulation to compare the proposed system with an ideal system (whole system control frequency of 10kHz) (Fig. 5). It is supposed that a haptic end-effector displays force against a virtual wall in simulation. We examined the force trajectory with changes in the coupling impedance and the 10kHz haptic interpolating parameters (IHR) according to Table 1.

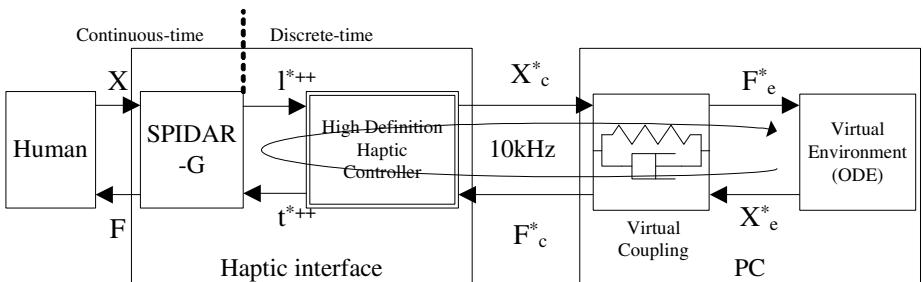


Fig. 5. System configuration of 10kHz ideal simulation

7.2 Simulation Results

In this simulation, the position of the haptic end-effector, as an input signal in this simulation, is shown in Fig. 6-(1), and the force trajectories A to C and D to F are shown in Fig. 6-(2) and Fig. 6-(3) respectively.

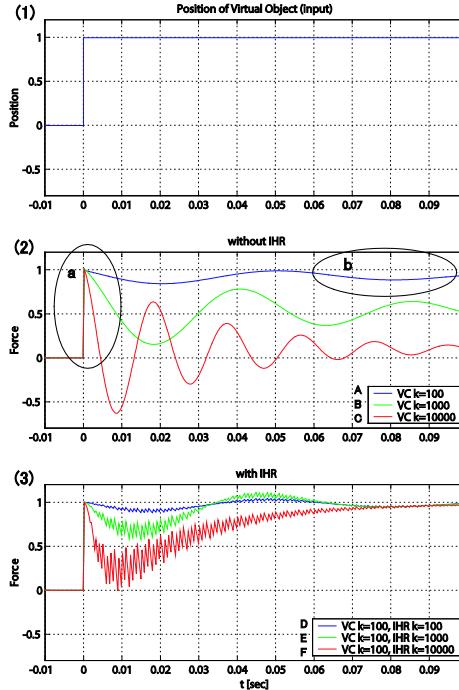
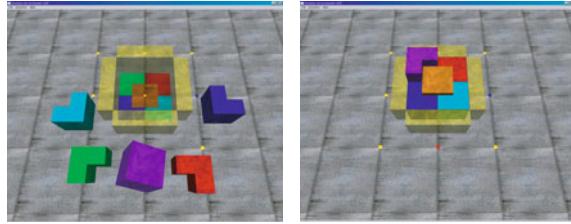
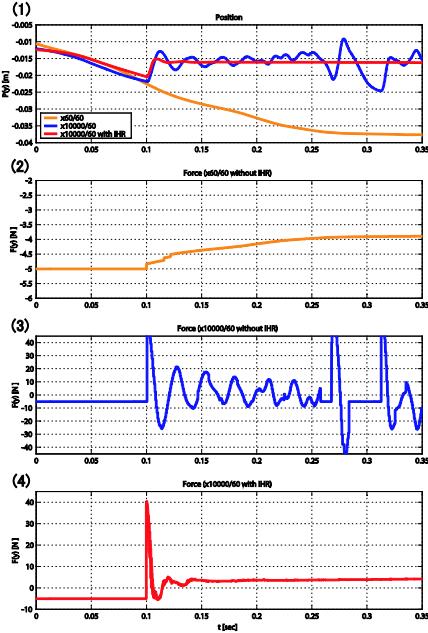
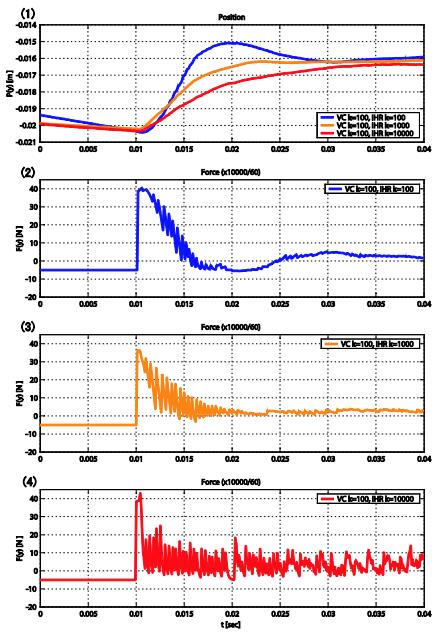


Fig. 6. Simulation Results

To compare the force in Fig. 6-(2) and (3), the trajectory is normalized to when the contact occurred. Fig. 6-(2) is the result of the ideal system. This result indicates that increasing a coupling impedance (K_c) makes the falling edge at the contact point (Fig. 6-(2)a) steep. This movement corresponds to the characteristics of the trajectory in [12] [2]. From the results of [12] [2], the stiffness of a real object becomes higher, and then the falling edge at contact point becomes steeper when we tap an object on the table in the real world. The trajectory after contact is restored to its normal state (Fig. 6-(2)b), depending on the coupling impedance (K_c). In comparison, Fig. 6-(3) shows the result of the proposed approach. This result indicates that in spite of the low coupling impedance, the IHR impedance was equal to that of a high coupling impedance at the contact point. At steady state, the trajectories with IHR were restored to the convergent trajectory of low coupling impedance. At the contact point, a haptic end-effector was coupled strongly to a virtual object, resulting in high fidelity. In the other manipulation of a virtual object, the coupling impedance became inherently low, creating high stability of manipulation. Therefore, this proposed

**Fig. 7.** 3D-puzzle**Fig. 8.** Result of displaying virtual wall**Fig. 9.** Result of displaying virtual wall (high impedance)

system achieved both stability and fidelity in manipulation of a rigid virtual object. There was some noise until steady state increasing the interpolation stiffness. The noise, which did not occur with high coupling impedance in the ideal system configuration, seemed to have been caused by an error of the interpolation. However, the noise is very effective in replicating hard surfaces, yet does not induce severe vibrations. In [12] [2], the noise is one of the most important factors in displaying a hard surface, when recording a pulse signal or effective vibration from a haptic end-effector, resulting in high fidelity.

7.3 Measurement of Displaying Force

We confirm stability and fidelity on the 3D-puzzle VR application(Fig.7). We measured position and force trajectories of tapping a virtual wall with a virtual object

connected to the haptic interface. Fig. 8-(2)(3) are the results of the traditional haptic system with only virtual coupling(2) and with virtual coupling and without IHR(3) respectively. Traditional haptic systems cannot achieve a sharp surface or a stable surface like real one. Fig. 8-(4), Fig. 9 are the results of the proposed system with 10kHz IHR. The characteristics of the force trajectory in Fig. 8-(4) are very close to the ideal result when tapping an object surface in the real world [12] [2]. These results showed that applying our proposed system eliminates the problems of stability and fidelity in displaying virtual objects.

8 Conclusion

In this study, we proposed and implemented a new system configuration which achieved both stability and fidelity in a video-rate VR application using the 10kHz high definition haptic rendering whose frequency ratio is approximately 160 times. We confirmed that our proposed system could achieve the stability and fidelity of haptic rendering through a computer simulation and 3D-puzzle application. Results showed that this system provided both stability, achieved by the virtual coupling with a low coupling impedance, and fidelity, achieved by the 10kHz high definition haptic rendering with a high interpolating impedance, in manipulation of a virtual object.

References

1. Adachi, Y., Kumano, T., Ogino, K.: Intermediate Representation for Stiff Virtual Objects. In: Proc. IEEE VR-AIS, pp. 203–210 (1995)
2. Okamura, A.M., Hage, M.W., Dennerlein, J.T., Cutkosky, M.R.: Improving Reality-Based Models for Vibration Feedback. In: Proc. of the ASME DSCD, vol. 69(2), pp. 1117–1124 (2000)
3. Baraff, D.: Analytical methods for dynamic simulation of non-penetrating rigid bodies. Computer Graphics 23 (1989)
4. Barbagli, Pratichizzo, Salisbury: Multirate analysis of haptic interaction stability with deformable objects. In: IEEE DC (2002)
5. Cavusoglu, Tendick: Multirate simulation for high fidelity haptic interaction with deformable object in virtual environments. In: ICRA (2000)
6. Colgate, J.E., Schnkel, G.: Factors affecting the z-width of a haptic display. In: Proc. of the IEEE ICRA, pp. 3205–3210 (1994)
7. Colgate, J.E., Stanley, M.C., Brown, J.M.: Issues in the haptic display of tool use. In: IEEE/RSJ ICIRS, Pittsburgh, PA, pp. 140–145 (1995)
8. Adams, R.J., Hannaford, B.: Control Law Design for Haptic Interfaces to Virtual Reality. IEEE TCST 10, 1–12 (2002)
9. Hasegawa, S., Ishii, M., Koike, Y., Sato, M.: Inter-Process Communicaton for Force Display of Dynamic Virtual World. In: Proc. of the ASME-DSCD, vol. 67, p. 11 (1999)
10. Kim, J., Sylvia, I., Ko, H., Sato, M.: Integration of Physics Based Simulation with Haptic Interfaces for VR Applications. In: HCI International (2005)
11. Akahane, K., Hasegawa, S., Koike, Y., Sato, M.: A Development of High Definition Haptic Controller. In: World Haptics, 3.18–3.20 (2005)
12. Kuchenbecker, K.J., Fiene, J., Niemeyer, G.: Improving Contact Realism through Event-Based Haptic Feedback. IEEE TVCG 12(2), 219–230 (2006)

13. Kawamura, S., Ito, K.: A New Type of Master Robot for Teleoperation Using A Radial Wire Drive System. In: Proc. IEEE/RSJ ICIRS, pp. 55–60 (1993)
14. Love, L., Book, W.: Contact Stability Analysis of Virtual Walls. In: Proc. Of DSCD, pp. 689–694 (1995)
15. Minsky, M., Ouh-Young, M., Steele, O., Brooks, F.P., Behensky, M.: Feeling and Seeing Issues in Force Display. Computer Graphics (ACM) 24(2), 235–243 (1990)
16. Novodex, <http://www.novodex.com/>
17. OpenDynamicsEngine, <http://opende.sourceforge.net/>
18. SPRING HEAD, <http://springhead.info/>
19. Terzopoulos, D., Platt, J.C., Barr, A.H.: Elastically deformable models. Computer Graphics (Proc. SIGGRAPH) 21, 205–214 (1987)
20. Zilles, C., Salisbury, J.K.: A constraintbased god-object method for haptic display. In: IEEE/RSJ ICIRS., HRICR., vol. 3, pp. 146–151 (1995)

Designing a Better Morning: A Study on Large Scale Touch Interface Design

Onur Asan, Mark Omernick, Dain Peer, and Enid Montague

Mechanical Engineering Building, 1513 University Avenue, Madison, Wisconsin, USA
{asan, omernick, dpeer, emontague}@wisc.edu

Abstract. In this paper, we described the design process of an individual prototype as it relates to Large Scale Public Touch Interface system design as a whole, and examine ergonomic and usability concerns for Large Scale Public Touch Interface (LSPTI) designs. The design process includes inspirational design, contextual design, storyboarding, paper prototyping, video prototyping and a user testing study. We examined the design process at each stage and proposed improvements for LSPTIs. Results indicate that the ‘color-field’ interaction methodology might be a good alternative to traditional ‘tabbed-hyperlink’ interaction in LSPTI implementations.

Keywords: Intelligent space, interactive design, touch screen.

1 Introduction

Technology advancement allows information to be accessible from more places, thus making our life easier and more efficient. From directories in shopping malls, to weather and news in school buildings, the use of intelligent space seems to fit smoothly into our daily life. Advancement in manufacturing has reduced price of touch screens making it more cost feasible for designers to develop a product based around a Large Scale Public Touch Interface (LSPTI). LSPTIs are primarily characterized in two ways: first, in scale; while monitor-sized touch-screen installations are already part of modern installations in places like museums, LSPTIs leverage the reduced cost of touch-screen technology to create a device that takes up a large percentage of the user’s field of vision (FOV). Second, the LSPTI is characterized by limited user interaction time. LSPTIs are used in indoor and outdoor settings, such as parks and museums[1]. Touch interaction allows the user to become more connected with the information he/she interacts upon, a new design dimension emerging in this era of graphic interaction [2]. Along with that, the high rate of touch screen phone use in the consumer market is a good sign of user receptiveness to interacting with Touch User Interface (TUI). Touch screen interfaces might also be a good alternative to traditional keyboards [3]. Therefore, LSPTI is a promising area awaiting more research and development. There are two design problems we would like to address in this study:

1. How should we present applications and textual data so that they would be usable by everyone?
2. What is the most efficient way to navigate through information on a Large Scale Touch Interface?

As there is still no standard set for the large scale TUI development, we would like to put our effort in providing ground-laying recommendation on best practice in designing such Interface. To do so, our team ran various usability tests on our prototype, a mockup of a LSPTI system designed for a home bathroom setting. The device as tested had three core functions: checking weather via an internally-designed interface, checking a calendar via the Mozilla Sunbird applications, and checking email via a trimmed-down browser application.

The proposed “Better Morning” mirror introduces an easy access information spot to a bathroom. Due to the nature of bathroom, the prototype is intended for shorter face-time. In the following paragraphs, we comment on elements that assisted us in the execution of our test plan.

Magical Mirror provides simulation of interface that integrates information accessibility and basic controls on a house appliance [4]. The project introduces us to the idea of home automation [5], which forces us to define the boundary we wish to design upon. How many applications would be sufficient for a touch screen interface and what are the design implications?

Smart Mirror demonstrated implementation steps and functionality of the interactive mirror [6]. The mirror was developed as an extension to home automation system that facilitates the integration of refreshment, appliances and various customized information services.

The goal in this study is to lay the groundwork for examining the effect of ergonomic concerns on user satisfaction with LSPTI implementations and examining other key factors to LSPTI user interface design. Effective LSPTI design needs to bring together ergonomic concern, visibility, and usability into an efficient interactive experience. Every touch should result in meaningful input to the device, so presenting data in the most easily understood format is crucial.

2 The Design Process

The first step in the design process involved every member of the team individually completing a write-up of inspirational designs in the world around us. We found problems that society is facing today, and then found several new inventions or innovations that are helping to solve these problems. This stage, while not directly related to the final design, gave the opportunity to really break things down and learn what exactly makes a design effective and usable.

From there, the next step was looking at the user’s perspective and completing a contextual inquiry. Here, we observed an eager participant performing some aspect of their daily life. This step was so useful in the design process, since it indicated that each people may perform a task or a routine in a certain way, the majority of the population often differs in their methods. This step laid the groundwork for later user-testing and stressed the importance of gathering useful data that will help improve the final design.

“A Better Morning” as we call it, is an interactive bathroom mirror that allows users to perform several parts of their typical morning routine at the same time. It is a large touch screen interface that allows a user to perform activities such as checking the weather, checking the daily calendar, or checking email, all on a bathroom mirror. The idea behind “A Better Morning” was that it would allow the user to multitask and take the time that he or she already spends in front of a mirror, while brushing his or her teeth for instance and maximizing the value of that time by allowing the user to do much more.

The next step was to create storyboards and identify key scenarios in which this product would be used. These storyboards allowed us to get down to details and show how the product fits into a user’s everyday routine, and how it will improve their quality of life.

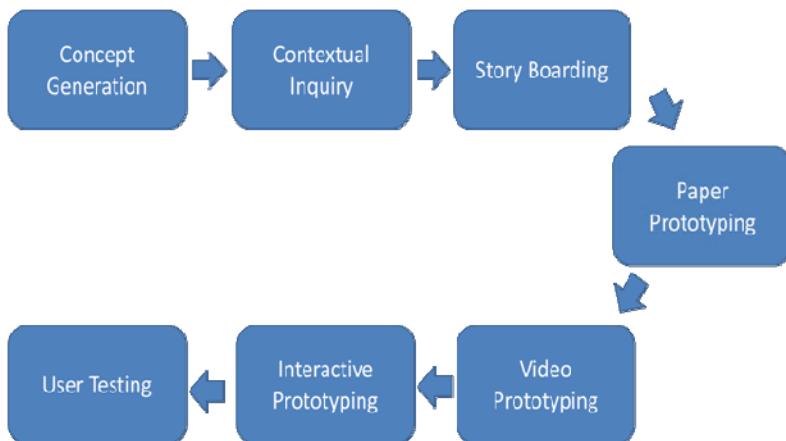


Fig. 1. Design process

2.1 Paper Prototype

The paper-prototype was the first real representation of the product design. We created a paper prototype that would allow participants to get a feeling for the system and interact with it in a very primitive way. This prototype was an easy way to allow us to see if our interface thus far matched a user’s mental model and how we could improve it. To collect this information, we outlined three tasks for users to perform while one group member ran the prototype and the others observed. These tasks included checking the weather, checking the daily schedule, checking three separate email inboxes and responding to an email. We then had a participant complete the tasks, as we observed and took notes. The most notable thing that we learned from this step was that the method to interact with the design must match the user’s mental models.

2.2 Video Prototype

The next step in improving the fidelity of the prototype was to create a video prototype depicting a single user scenario. This prototype was more scripted than the paper prototype and showed less interactivity, but it did a good job of showing how the product fits into a user's daily routine, and how quality of life would be improved with this system. It portrayed a single user's morning routine without the system, and the much more streamlined routine that is made possible with the system. The video prototype allowed us to see the system in much higher fidelity and also highlighted some of the key limitations that would define the rest of the design process. We found that because of the uniqueness of our product concept, prototyping it in true form was a nontrivial task and was something that was out of the scope of the project given time constraints. We were going to have to step back and get creative to develop a way to truly test the system's interactivity in the next stage.



Fig. 2. Screenshot from a Better Morning Video Prototype

2.2 Interactive Prototype

The final step, and first real incarnation of the "A Better Morning" was the interactive prototype. This prototype consisted of an interactive PowerPoint representation of the system, projected from behind users onto a surface to mimic an actual mirror. The PowerPoint representation had all of the functionality needed to complete the tasks as previously outlined in the testing of the paper prototype. The most useful part of this prototype was the user testing that it enabled us to do. Six participants completed these tasks with the interactive prototype while we gathered data from entrance interviews, direct observations and exit interviews. This prototype was as close as we could possibly get to the actual system without physically building it. It was one of the most useful steps in the design process to this point, because of the volume and quality of feedback that were collected.



Fig. 3. Screenshot from video of A Better Morning Interactive Prototype user testing

3 User Testing

The user testing was conducted in a human computer interaction lab. We used a projector to act as the smart mirror on the wall. The system was established as PowerPoint, and one of the testers controlled it. When participants clicked something on the wall, it was clicked by the tester at the computer. The environment was not a real bathroom environment, but there was a preparation session for participants and they were asked to role play.

3.1 Participants

There were total 6 participants, three of them were female. They were all students at UW-Madison. All of the participants were frequent computer users. They all declared what kind of activities they do in the morning such as spending a significant amount of time at bathroom. Five of the participants told us that they check internet, emails or news as morning activities. We also measured the arm length and height of the participants.

Table 1. All demographic information about the participants

Arm Length	#	Height	#	Hand	#
25"-29"	3	5'-5'	2	Right	5
30"-35"	3	5'6"-6'	2	Left	1
		6'1"-6'5"	2		

3.2 Procedure

The user testing consisted of a short pre-interview, a quantitative user observation session, and a short post-interview. During the pre-interview, one of the testers described the functionality and aim of the prototype device and collected physical

user data (arm length and height) and qualitative data related to their routine morning activities. The user observation session consisted of three tasks, which were divided by ‘simple’ (requiring a minimum of $i \leq 3$ interface interactions) and ‘complex’ (requiring a minimum of $i \geq 4$ interactions). These tasks are:

- 1) Check today’s weather. (Simple, $i=1$)
- 2) Check the calendar. Examine the detail of Nov 4, which has a meeting option. Print out the detail. (Complex, $i=4$)
- 3) Check all of the email accounts linked to the system and respond to the email in the Gmail account. (Complex, $i=6$)

During the study, all participants were observed to collect qualitative data. These data are errors made per task, numbers of interface touches per task, completion time per task and task completion success or failure. In the post-interview, participants were asked to give feedback about the prototype. There were questions about rating their satisfaction level, ease of use for the prototype, simplicity and responsiveness of the prototype as well as their improvement suggestions for the system.

4 Result

All participants completed all three tasks successfully. Task completion time varies for each task and for each participant. For task 1 the average time was 14 seconds, for task 2, 35 seconds, and for task 3, 53 seconds.

We were also interested in number of screen touches used to complete each task. The following averages was obtained; task 1 ($M=2$), task 2 ($M=6$), task 3 ($M=9$). Another metric that was observed was the number of errors the participant made during the tasks. The error was defined as “any interaction with the prototype that took the user further from the interaction goal”. The following numbers for each task: was obtained; task 1 ($M=0.33$), task 2 ($M=1$), task 3 ($M=1.66$).

After completing all tasks, participants found the product easy to use (4.83 on a scale of 1 to 5) and simple to use (4.5 on a scale of 1 to 5). The satisfaction level of participants was just below moderately satisfied (3.67 on a scale of 1 to 5).

4.1 Usability vs. Physical Attributes

One of the key aspects of the system examined in this prototyping session was the correlation between physical attributes (height, arm length) and core satisfaction, which was measured by the first two questions in the post-user test. The results indicate that there seems to be a correlation between height and ease of use – the only individual who did not score the system 5/5 for ease of use was the shortest individual in the test study, and she further noted that the system was ‘difficult to reach’. Based on this, it is possible to infer a similar correlation between ease of use and arm length, which contains a similar graph. These results indicate that the layout of the system interface presents usage challenges for individuals under 5’.

Despite the above result, there seems to be little, if any, correlation between either height or arm length and overall satisfaction, as seen below.

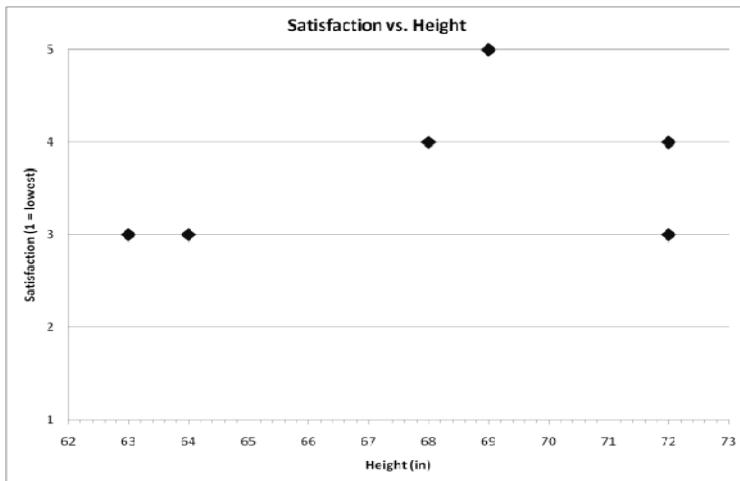


Fig. 4. Satisfaction vs. Height

These data imply other limiting factors for overall satisfaction. An examination of the exit interview free-response section reveals that 66% ($M=4$) of users commented negatively on the size of the fonts used in the system. Furthermore, during testing, several users indicated that they could not read the fonts used. However, the smallest font used was approximately twice the size of a standard 12 pt. font read at the same distance. Since a 12-point font should be very readable ,^[4, 5] another factor must be involved.

Due to testing limitations, the projector we used was low-resolution, causing each letter to have a height of approximately 4 pixels. In addition, no antialiasing was available. Since decreasing legibility is very strongly correlated with both lack of antialiasing and decreasing the number of pixels in the letter-shape [5], the results indicate that it is not the size of the text that was the limiting factor, but rather the resolution of the display.

4.2 Design of a Simple Interface

Another key aspect of the system examined during this prototyping session was the simplicity of the interface. Interface simplicity was a key design goal throughout the prototyping process, and the prototype scored relatively high, marking a 4.5/5 with no measured correlations. This suggests that the prototype interface represents a useful foundation for further expansion. However, the data gathered during the test has implications for the design paradigm used in LSPTI implementations.

The interface testing revealed major differences between error rate on each task. Here, error rate is defined as a user interface reaction that moves the user away from the task goal. First two tasks were easily completed, with an average of 0.5 interface errors per task. The third task, however, had significantly more errors, with an average of 1.667 interface errors per task. This suggests that the design of the interface used in the third task is significantly less efficient than those used for the first two

The first and second interfaces are based around ‘color-field’ interactions, where most interactive zones are represented by either large fields of color or icons. The third interface, however, is based around ‘tabbed-hypertext’ interactions designed to be reminiscent of popular Internet browsers. The results suggest that the ‘tabbed-hypertext’ design paradigm made users expect a certain set of interactions with the interface, and that the interface did not align with these expectations.

The interface testing also revealed some shortcomings in the application of the ‘color-field’ style interface. In the interface used for the first task, we used color fields to delineate non-interactive data areas in the display in addition to interactive zones. In the interface used in the second task, all color field areas were interactive zones. This difference resulted in a large difference in user deviation from optimal interaction paths as measured by the ‘optimality ratio’, or the ratio of non-optimal interactions per optimal interaction. In the first task the average optimality ratio was .83, while in the second task the average optimality ratio was .42, despite the second task requiring more interactions to successfully complete. This indicates that the second interface, where color fields are used exclusively to delineate interactive zones, represents a more intuitive user interface.

Table 2. Task 1, a simple task (i=1) using color-field interface

Completion time (sec.)		# of Touches	# of Non-optimal touches	# of Errors
Mean	14.33	1.83	0.83	0.5
S.D	8.80	0.75	0.75	0.55

Table 3. Task 2, a complex task (i=4) using color-field interface

Completion time (sec.)		# of Touches	# of Non-optimal touches	# of Errors
Mean	35.33	5.67	1.67	0.5
S.D	6.62	1.03	1.03	0.55

Table 4. Task 3, a complex task (i=6) using a tabbed-hyperlinked interface

Completion time (sec.)		# of Touches	# of Non-optimal touches	# of Errors
Mean	55.67	9.33	3.33	1.7
S.D	13.85	1.37	1.37	0.82

5 Discussion

This study has revealed several areas for improvement in the design of LSPTI. First, we found that an LSPTI should be able to adjust to the reach ability of the user, in

order to avoid placing interactive zones outside of the user's reachable area. While this would not normally cause a concern for smaller TUI applications, the inherent larger size of LSPTI installations requires this adjustment. The results suggest that this is one of the main causes of decreased ease of use. To accomplish this, either some sort of interface control could be created that would allow a user to 'drag' the interface down to a reachable level, or perhaps some sort of spatial-recognition feature could be integrated into LSPTI-based devices which would be able to ascertain the rough physical dimensions of the user.

Second, we found that high-resolution displays are very important for future LSPTI implementation. While a low-resolution projector may be appropriate for presenting at a distance, they are not as useful when a user is within arm's length of a screen. At 20-30 range, the amount of data that can be discerned on a large, low-resolution device is very low, and can cause user errors and dissatisfaction.

Third, we found that the current model of 'tabbed-hypertext' browser interactions does not serve the design of LSPTI well, allowing more user errors than the proposed 'color-field' interaction methodology. 'Color-field' takes better advantage of the larger space available on LSPTI implementation and allows users to interact with larger active areas, ultimately reducing the number of interface errors. Given that LSPTI installations may be found in areas where the user would have a limited ability to use the interface to the point where he or she could develop a strong mental model, reducing interface complexity is the key factor.

Finally, we found a key limitation of the 'color-field' interaction methodology. If a designer uses 'color-field' in a LSPTI implementation, there must be a clear distinction between which areas are interactive zones and which areas are non-interactive, used only to delineate different data areas. Failure to do this will lead to user confusion and inefficient interface traversal.

6 Future Study

There is a great deal of future work to be done in the LSPTI field. The wide range of additional applications suggested by our participants indicates that LSPTI-based interfaces have a wide potential for further development. The preliminary results encourage further studies in the areas of 'color-field' interaction methodology as an alternative to traditional 'tabbed-hyperlink' interactions in LSPTI implementations.

References

1. Nakanishi, H., Koizumi, S., Ishida, T., Ito, H.: Transcendent communication: location-based guidance for large-scale public spaces. In: Conference Transcendent Communication: Location-based Guidance for Large-scale Public Spaces, p. 662. ACM, New York (2004)
2. Panchanathan, S., Kahol, K.: Guest Editors' Introduction: Haptic User Interfaces for Multimedia Systems. In: IEEE Multimedia, pp. 22–23 (2006)

3. Plaisant, C., Sears, A.: Touchscreen interfaces for alphanumeric data entry. In: Conference Touchscreen Interfaces for Alphanumeric Data Entry, pp. 293–297. Human Factors and Ergonomics Society (1992)
4. Tinker, M.: Legibility of Print for Children in the Upper Grades. American Journal of Optometry and Archives of American Academy of Optometry 40, 614 (1963)
5. Sheedy, J., Subbaram, M., Zimmerman, A., Hayes, J.: Text legibility and the letter superiority effect. Human Factors 47, 797 (2005)
6. Hossain, M., Atrey, P., El Saddik, A.: Smart mirror for ambient home environment. In: Conference Smart Mirror for Ambient Home Environment, pp. 24–25 (2007)

Experimental Evaluations of Touch Interaction Considering Automotive Requirements

Andreas Haslbeck, Severina Popova, Michael Krause, Katrina Pecot,
Jürgen Mayer, and Klaus Bengler

Institute of Ergonomics, Technische Universität München,
Boltzmannstrasse 15, 85747 Garching, Germany

{Haslbeck,bengler}@tum.de, {popova,Krause}@lfe.mw.tum.de,
kjpecot@gmail.com, Juergen.Mayer@eu-bs.de

Abstract. Three different usability studies present evaluation methods for cross-domain human-computer-interaction. The first study compares different input devices like touch screen, turn-push-controller or handwriting recognition under regard of human error probability, input speed and subjective usability assessment. The other experiments had a focus on typical automotive issues: interruptibility and the influence of oscillations of the cockpit on the interaction.

Keywords: Touch, input medium, efficiency, effectiveness.

1 Introduction

The evaluation of different input devices is a classical question in usability experiments. Meanwhile a multiplicity of alternatives for most different employment scenarios exists. Nowadays such input devices are not only taken for human-computer-interaction (HCI) at PC workstations, but they are spread over many different professional applications, like cars, machinery, industrial workplaces and also taken for leisure and mobile applications. A recent example is Apple'stm launch of the iPhone: touch-based interaction became popular and so it now emerges in more and more applications especially at mobile devices and in new domains. Above all, many new gadgets address user experience and an emotional channel of its potential users.

One focus on these trends is automotive driven: which input device is the best medium for driver's interaction with In-vehicle Information Systems (IVIS). Within this domain, different use-cases exist, beginning from cars to heavy trucks and construction machinery. All these scenarios have different requirements on usability.

This paper presents three experiments, which focus on effectiveness and efficiency of different input devices depending on the context where they are to be taken in use. The focus is on the one hand on typical operating steps towards IVIS and on the other hand on different environmental conditions which are typical for an automotive application. The different experimental settings are also recommendations for automotive HCI designers to evaluate future interfaces in early stages of product development.

2 Present Situation

All devices and use-cases have in common that their usability strongly depends on the man-machine-interface (MMI). The most important mean of communication is a hand-based interaction via touch media or different turn-push-controllers. Past alternatives are mechanical hardware buttons, levers and rotating discs, while speech recognition has in fact been implemented, but is rather a future trend and has not yet reached full performance.

A previous study, by Kavakli and Thorne [1] compared the user's performance on computer video games using different input devices. It revealed that the kind of game has an influence on the best input device to take in order to improve user's performance. Nowadays another important focus lies on the age of users: the question is whether the user's age has an influence on operator's performance with different input devices. Murata and Iwase [2] showed that users of different age groups revealed no significant differences in performance at a pointing time experiment on touch panels, while they did on the same setting, operating with a computer mouse. In addition, they also stated that learning time required by a touch panel was smaller. In contrary to these findings, an experiment conducted by Armbrüster et al [3] found differences depending on the age while different tasks (click and drag & drop) had to be done via touchpad. Rogers et al [4] concluded in studies similar to the following experiment, that younger users are able to show faster input times than older ones.

3 Alphanumeric Input

In a first experimental setting, different input media have been examined during alphanumeric inputs concerning effectiveness, efficiency and satisfaction in a single-task situation. The participants had to write down texts under the demand of a minimum of faults and a maximum of typed characters. So the subjects had to handle a speed-accuracy-tradeoff. Among the different input devices, a traditional keyboard, a touch screen keyboard on laptop and iPod Touch, a knob like it is equipped in cars and a handwriting recognition software on laptop were taken into account.

3.1 Participants

The test subjects have been university employees and students. 15 participants between the ages of 24 and 60 (mean = 31.7, SD = 11.1) took part in this study. There were a total of 12 males and 3 females. All the participants had normal or corrected to normal visual acuity and gave their consent for the experiment. Participation was done on a voluntary basis.

3.2 Experimental Design and Methods

The subjects were required to use five different devices for text input. These were a Lenovo S10-3T touch screen (TS) tablet PC featuring a 10 inch display with two text input options (touch-keyboard and handwriting recognition), a traditional Dell QWERTY keyboard, a knob-like device reminiscent of the BMW iDrive (see fig. 3)

but with different mechanical qualities and in order to have comparison to a leisure gadget, an Apple iPod Touch (see fig. 4). With the knob device, the users scrolled through a wheel of letters and pressed down on the knob to select the intended letter. While handwriting, the user wrote by finger on the screen and the computer software recognized the writing and converted it to text. The hardware keyboard and the knob were attached to the laptop so all alphanumerical inputs could be done by the laptop with the exception of the iPod. While writing on the iPod the decision to hold the iPod upright or crosswise was left to the participants.



Fig. 1. Touch keyboard



Fig. 2. Handwriting recognition form



Fig. 3. Tentative turn-push-controller



Fig. 4. iPod Touch

In the setup of the experimental apparatus, the tablet-PC was located in front of the participant at a comfortable viewing distance. It was placed on a wedge at a 45 degree angle to allow viewing the laptop in an angle of nearly 90 degrees. An additional LCD screen was present right in front of the subjects, which displayed the alphanumeric text that the user was asked to copy in each trial with the present device. In total, there were five different alpha-numeric input texts. The texts consisted of simple sentences from sources such as newspapers and books, containing several numbers, for example short news from a football match. The order of administration of the device and text sentences was randomized for each user.

With exception of the iPod, the text input progress was observed for each device by the test administrator on another LCD screen that duplicated the laptop screen.

The research question of this experiment is, whether the kind of input device has an influence on human error probability (HEP) and input speed. The subjects had three minutes to write on each device after the chance of getting familiarized with all

devices. The instructor noted the exact word/number when passing the first and the second minute of each trial in order to have information about writing progress every minute. The assessment of the user's performance at the TS handwriting recognition has limits: a typing error, which is common for keyboards, does not exist there. The reason is that every input is processed by the software and afterwards interpreted as an expression into the set language. On account of this, here the counted errors are mistaken words by the resulting inputs after the recognition software process. So many counted errors have its origin in spidery handwriting, but less in problems of typing or spelling words or the design of this input device. The subjects were also told not to care for errors while typing in order to waste no time.

For subjective measurement of the experiment, participants completed the System Usability Survey (SUS) by Brooke [5]. The SUS is a simple, ten-item Likert scale which gives a global view of the subjective assessment of the usability of a device.

Objective measurement has been done by text analysis after the experiments. For each input medium, the average input speed was calculated in characters per minute. This was done by counting the total characters and spaces typed with each device per participant, then dividing the total characters by three and averaging them for each medium. Looking at table 1, it can be seen that the keyboard had the fastest input. The TS keyboard, TS handwriting recognition and iPod had relatively similar input speeds, and the knob had the slowest input speed, which, in comparison to the keyboard, only had about 10% of whose input speed.

The HEP, which is defined as the ratio between the number of actual occurred errors and the number of given opportunities for errors to occur, was calculated for each input device. To obtain this number, the total errors and characters were each summed across the participants on each input medium. Then they were divided (total errors across participants / total characters performed on the medium across participants) to obtain the HEP. In table 1 can be seen that the keyboard had the lowest error probability. The iPod had the second lowest HEP. The remaining input mediums, TS handwriting, TS keyboard and knob, had the highest HEPs.

Finally a correlation analysis was performed on the data for each input medium to determine if there was a correlation between the speed of input and errors performed. From these analyses, two input devices (the hardware keyboard and TS keyboard) showed strong positive correlations of $r_p = .702$ with $p = .004$ (hardware keyboard) and $r_p = .731$ with $p = .002$ (TS keyboard). Therefore, as the speed of input increased, the number of errors increased. The other devices showed no significantly strong correlation values.

These results show the advantages of a hardware keyboard in comparison to all other input devices: it enables a high input speed combined with low error rates. So this is the strong recommendation for workplaces where this type of keyboard is applicable. In spite of these facts, the user's choice can be different: in leisure application the mentioned performance measures don't count. Here, look and hedonic aspects are crucial. An additional AttrakDiff [6] evaluation of the different input devices has shown, that in terms of hedonic quality, the handwriting recognition, the iPod Touch and the TS keyboard have gained very good results, much better than the hardware keyboard. And finally, these results are valid only for single-task situations in a static environment.

Table 1. Results of alphanumeric input experiment

	average char./min	standard error char./min	average HEP	standard error HEP	average SUS Score	standard error SUS score
hardware keyboard	201.1	11.7	0.020	0.003	87.3	3.6
TS keyboard	81.3	4.9	0.039	0.006	68.5	4.2
TS handwriting rec.	61.9	3.7	0.034	0.004	62.7	3.8
iPod touch keyboard	72.0	5.6	0.028	0.009	66.7	4.7
turn-push-controller	23.6	1.3	0.038	0.004	56.0	4.2

4 Touch vs. Knob under Regard of Interruptibility and Oscillation

The following two experiments concentrate on a touch screen and a turn-push-controller under two different experimental conditions: interruptibility, tested by the occlusion method and oscillation simulated by a hexapod.

4.1 Experimental Tasks

For both of the following experiments, the tasks for the participants were the same. As a preliminary step, a function typology was developed from an exemplary list of functions for IVIS. This includes typical functions that appear in human-computer-interaction in the motor vehicle. From the functional typology, individual interaction types were derived systematically. For example, alpha-numeric input can be described as a sequence of the interaction types "selection from a list" or "select one from n elements" and "confirmation". With this technique any kind of typical in-vehicle interaction can be described as a series of these basic interaction types.

The following interaction tasks have been used for experimental research: select one from n elements (fig. 5 left), select one element from a list, two-dimensional tracking task (fig. 5 center), setting of a certain value (fig. 5 right) and confirmation by pressing the knob or touching the screen. For the selection of one element from n elements the reaction time was recorded from the moment when the elements occur until the moment when the correct (or incorrect) square was pressed. In the touch screen condition, the intended element was to be touched directly, while in the knob condition a pointer switched with each step turning the knob from left to right and went down one line when the present line was crossed completely. The same time measurement was done for the selection from a list task and the setting of certain given values. For both tasks in the touch condition, a direct selection of the targeted

elements was possible again. In the knob condition, a pointer was directed down the list of possible elements, selection was done by pressing the knob. For the setting of a given value, the turning of the knob raised the values by turning right / clockwise and decreased numbers by turning left / counter-clockwise.

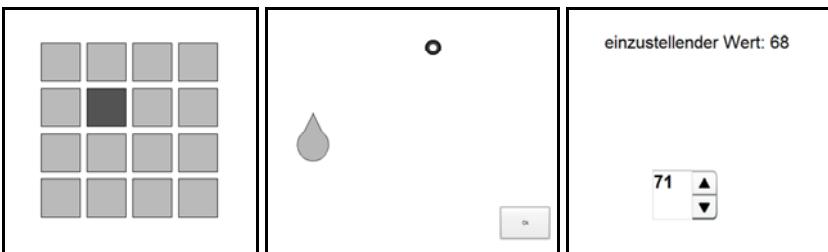


Fig. 5. Interaction prototypes: select one element from n elements, two-dimensional tracking, setting of a certain value

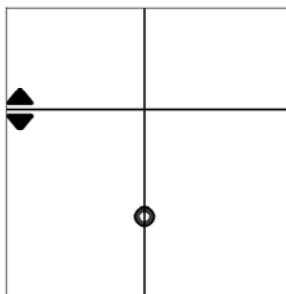


Fig. 6. Two-dimensional tracking with crosshairs in knob condition

For the two-dimensional tracking task the time needed to bring the pointer to the circle by drag & drop (touch condition) was measured. The measurement was triggered by touching the element. With the knob the two lines of crosshairs were moved one after the other (fig. 6). The position of each line was confirmed by pressing the knob.

4.2 Occlusion Experiment

The second experimental setting concentrated on a comparison between touch-based interaction and turn-push-controller: here, both have been applied on a set of individual basic operations. Nowadays, these two input devices are commonly equipped in high class cars, so the usability of both is in interest.

In this experiment the above mentioned tasks have been used. They were done under regard of operating speed and error rates. In addition, the interruptibility of the tasks was measured by employment of the occlusion method. One characteristic of the following experiment is the fact that the research question aims to the automotive use of both devices, but the tests are done in a usability lab in an early stage of product development and no real car or driving simulation is needed.

Participants. 20 participants took part in this study. The 15 male and 5 female test subjects have been university students and employees. The age was between 22 and 54 with a mean age of 29.3 years ($SD = 6.8$). All participants were right-handed, two of them had red-green color blindness.

Experimental Design and Methods. After familiarizing with both input devices, all operating tasks have been conducted with and without occlusion conditions in a permuted sequence. According to ISO [7], the shutter-glasses (Plato Visual Occlusion Spectacles by Translucent Technologies Inc.) were set to occlusion and vision intervals of 1500 ms each. This method supports to simulate typical interruptions of the primary driving task by a secondary IVIS operating task. Nevertheless this experiment was a single-task operation. The specific metrics of this method, the total task time (TTT) and the total shutter open time (TSOT) are set into relation (R coefficient):

$$R = TSOT / TTT$$

TTT is to be measured in a baseline setting without occlusion. It is the time needed for fulfillment of the whole task, while in comparison the TSOT comes from the same experiment under occlusion conditions and therefore all time with open glasses is measured. The R coefficient expresses the interruptibility of an operating task. If R is smaller than 1, the task easily can be interrupted, while values larger than 1 characterize tasks which need much visual attention.

Beside the above mentioned objective measurements, subjective evaluation was done by usability questionnaires according to ISO 9241-11 [8].

Results. During the experiment, the single time of every operating step, the operating time for the whole tasks and the operating errors were measured. Here the participants had to do a speed-accuracy-tradeoff. The results show again that a suitable use of both evaluated devices strongly depends on the circumstances where they are applied. While being taken as a single task, the touch screen operation was faster than the usage of the knob. A further research question would be, if this result is also valid for secondary tasks. It might be possible, that the input on a knob could be more reliable, because of the interruptibility. A comparison of the occlusion-related R values showed, that interaction with the turn-push-controller can rather be interrupted:

$$R_{\text{turn-push-controller}} < R_{\text{touchscreen}}.$$

The distance of the touch spot to the target center point for different parameterized sections of the two-dimensional tracking task was calculated (with a bin width of 2 pixels; and an upper limit category of ‘distance ≥ 36 pixels’). The resulting figure of cumulative probability function (fig. 7) shows for example, that a target with 15 pixels radius (6.6 mm diameter) will be hit right by about 60 % of participants.

A closer look at the different experiments revealed, which interaction types are best fitted for the application of touch screen use in cars. We recommend to use icon menus in touch screen applications. Lists should be limited to a few entries, which fit to one screen. To enter distinct values the use of spin buttons or low resolving sliders should be favored. Two-dimensional inputs like in navigation tasks, can be easily achieved by positioning a drag & drop marker.

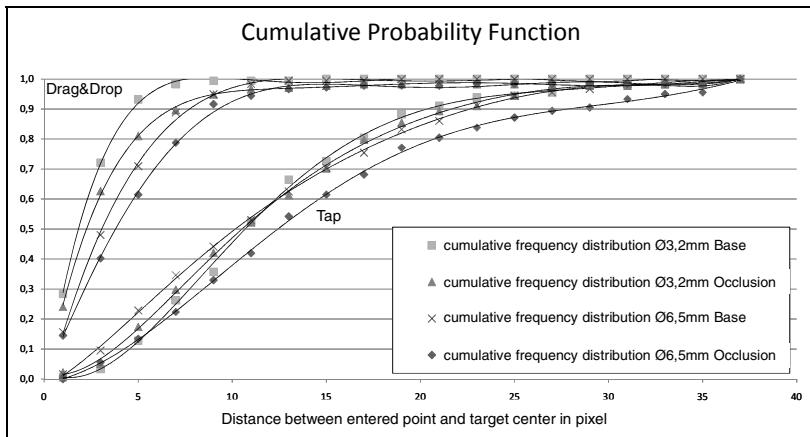


Fig. 7. Cumulative probability function

4.3 Touch and Knob under the Influence of Oscillations

Vibrations have an influence on human operators. Their impact can cause reduced working performance or restrain the fulfillment of tasks. A third experiment is to be introduced here as proposal for an experimental design. It focuses on the above mentioned devices, knob and touch screen, in combination with an oscillating environment. Here, both input media have been tested under vibration, which simulates a truck ride on a highway. Therefore a Steward platform (motion hexapod, six degrees of freedom, see fig. 8) was deployed with a sitting mock-up mounted on top (fig. 9). In this experiment the same tasks like in the second experiment have been conducted by the participants in a single task condition.

Participants. The 18 male participants, all from an academic environment, had an age between 25 and 64 with a mean age of 30.3 years ($SD = 10.4$). They were all right-handed and had normal or corrected to normal visual acuity. Two persons had red-green color blindness.

Experimental Design and Methods. In the last study a roadway-induced vibration was simulated to find out how this condition affects the operation with both input devices. The Steward platform was programmed with an oscillation of a truck driving on a German highway.

In the experimental procedure, the participants had to sit on a typical truck seat with fastened seat belts. The seat was mounted on the experimental mock-up. For safety reasons there was an emergency stop button mounted in a reachable distance to the participants which would stop all oscillations immediately. In no experiment this button was pressed. The movements of the hexapod were rather slight (according to a trip on a highway) so the subjects were able to concentrate on their operating tasks.

Here the participants had the following tasks: select one from n elements, two-dimensional tracking task, select one element from a list, setting of a certain value. All tasks have been conducted with and without oscillation conditions.



Fig. 8. Stewart platform for driving oscillation

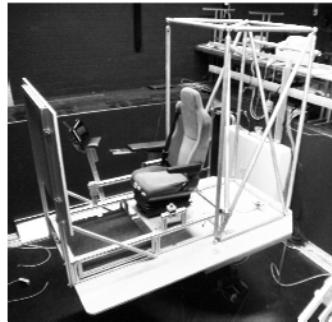


Fig. 9. Experimental mock-up

Results. With oscillation the results are very similar to the second experiment: the touch screen allows higher input speeds in the most tasks. In the list selection task with a long list (14 elements on two screens) and by setting of a certain value the knob achieves shorter input times and less input errors. The touch screen produced less input errors in the task “select one of n elements” when the elements had a size according to VDI/VDE 3850 [9]. In the two-dimensional tracking task the touch screen achieved shorter input times and a lower number of errors. These results correspond with most findings of Lin et al [10]. In contrast to this experiment, their oscillation simulated maritime conditions.

5 Conclusion and Discussion

For alphanumeric input, a traditional keyboard showed its strength in the first mentioned study (static condition). It allows a high input speed in combination with a low error rate. If such a device cannot be taken, a touch-based handwriting recognition has a high performance potential, if the recognition itself can be improved more and the recognition process reduces its vulnerability for spidery handwriting. A larger area for the script input could also be an improvement. The other experiments also showed important criteria for the HCI towards IVIS: when operating in dual-task settings, interruptibility and environmental influences have to be mastered. By the use of the occlusion method it was shown that a turn-push-controller can support a user in a frequently interrupted task more than a touch screen can do although the latter allows faster operating processes while being taken for single tasks – a finding which was also valid under the influence of moderate oscillations. The results also shed light on necessary sizes of elements to be pressed or reached in a touch screen based operations.

The discussed experiments have shown performance data of HCI with the evaluated input devices. But they also introduced some experimental settings and methods which are well established in lab investigations in order to bring them into new application domains. The spread of innovative input devices and strategies will always bring new ideas to the automotive domain. Next developments could be multi-touch gestures.

Yee [11] has discussed this kind of interaction and showed limitations. So further experiments could address multi-touch gestured under regard of interruptibility and oscillations in comparison to classical single-touch interaction.

Acknowledgements. The authors gratefully acknowledge the support of the TUM Graduate School at Technische Universität München, Germany.

References

1. Kavakli, M., Thorne, J.R.: A Usability Study of Input Devices on Measuring User Performance in Computer Games. School of Information Technology, Charles Sturt University, Bathurst (2002)
2. Murata, A., Iwase, H.: Usability of Touch-Panel Interfaces for Older Adults. *Human Factors* 47(4), 767–776 (2005)
3. Armbrüster, C., Sutter, C., Ziefle, M.: Notebook input devices put to the age test: the usability of track point and touchpad for middle-aged adults. *Ergonomics* 50(3), 426–445 (2007)
4. Rogers, W.A., Fisk, A.D., McLaughlin, A.C., Pak, R.: Touch a Screen or Turn a Knob: Choosing the Best Device for the Job. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 47(2), 271–288 (2005)
5. Brooke, J.: SUS - A quick and dirty usability scale. In: Jordan, P.W., Thomas, B., Weerdemeester, B.A., McClelland, I.L. (eds.) *Usability evaluation in industry*, Taylor & Francis, London (1996)
6. Hassenzahl, M., Burmester, M., Koller, F.: AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In: Szwilus, G., Ziegler, J. (eds.) *Mensch & Computer 2003: Interaktion in Bewegung*, 7. bis 10, September 2003, Teubner, Stuttgart (2003)
7. ISO, 16673. Road vehicles - Ergonomic aspects of transport information and control systems - Occlusin method to assess visual distraction due to the use of in-vehicle systems (2007): ISO
8. ISO, 9241-11. Ergonomic requirements for office work with visual display terminals (VDTs) - Part 11: Guidance on usability (1999): ISO
9. VDI/VDE, 3850. User-friendly design of useware for machines. Beuth Verlag, Berlin(2000)
10. Lin, C.J., Liu, C.N., Chao, C.J., Chen, H.J.: The performance of computer input devices in a vibration environment. *Ergonomics* 53(4), 478–490 (2010)
11. Yee, W.: Potential Limitations of Multi-touch Gesture Vocabulary: Differentiation, Adoption, Fatigue. In: Jacko, J.A. (ed.) *Proceedings of 13th International Conference on Human-Computer Interaction. HCI International 2009, Part II*, San Diego, CA, USA, July 19-24, 2009, pp. 291–300. Springer, New York (2009)

More than Speed? An Empirical Study of Touchscreens and Body Awareness on an Object Manipulation Task

Rachelle Kristof Hippler¹, Dale S. Klopfer³, Laura Marie Leventhal²,
G. Michael Poor², Brandi A. Klein³, and Samuel D. Jaffee³

¹ Applied Science - Firelands

² Computer Science Department

³ Psychology Department

Bowling Green State University, Bowling Green, OH 43403

{rkristo, klopfer, leventha, gmp, brandik, jaffees}@bgsu.edu

Abstract. Touchscreen interfaces do more than allow users to execute speedy interactions. Three interfaces (touchscreen, mouse-drag, on-screen button) were used in the service of performing an object manipulation task. Results showed that planning time was shortest with touch screens, that touchscreens allowed high action knowledge users to perform the task more efficiently, and that only with touchscreens was the ability to rotate the object the same across all axes of rotation. The concept of closeness is introduced to explain the potential advantages of touchscreen interfaces.

Keywords: Touchscreens, Reality based Interface Model, Cube Comparison Task, Mental Rotation in Virtual Environments.

1 Introduction and Background

Touchscreens are a popular interaction style for handheld devices. Early touchscreen research found that touchscreen interactions were faster than mouse interactions [15]. Yet contemporary advertising for touchscreens emphasizes object manipulation on the screen rather than speed *per se*; for example, users of phones with touchscreens are often shown rotating and moving photographs and other objects on the screen. Can touchscreens provide more than speeded interactions with virtual objects?

Following [6], we believe that interfaces in virtual environments (VE) should be designed so as to “allow participants to observe and interact within the VE *as naturally as they would interact with objects in the real world* [p. 84, italics ours].” What this means is that in the design of an interface, it is necessary to consider the user’s implicit knowledge of how to move his or her body and limbs; of how to navigate through, and manipulate objects within, the environment; and of how the laws of physics operate, if only at a low level of understanding. These three elements of interface design have been nicely captured by the *reality based interaction* (RBI) framework [7], and illustrate three of the RBI themes: body awareness, environmental awareness, and naïve physics.

When the user cannot interact with virtual objects as he or she would do so naturally, two other factors need to be considered. The first is a cognitive one: understanding of the result of a particular interaction with a device. Unlike procedural knowledge, which is often represented as condition-action pairs, *action knowledge* can be represented as action-condition pairs of the form “if I perform action X, Y will result.” Whether the result of a particular action is what the user intended is another, related matter. A second factor of interest in less-than-natural interfaces is *closeness*, or the degree to which an action performed via the interface differs from the action performed to achieve the desired result in the natural world.

Closeness and action knowledge are inversely related: as closeness increases, cognitive demands and action knowledge of the user decreases. In other words, as the interface becomes less natural, what the user needs to know both about how to perform some action and about what results from performing that action increases. Therefore, by avoiding the increased cognitive supplementation required by far interfaces, close interfaces should be easier to use (and to learn to use) than far ones (cf. [14]). In this paper, we elaborate on the RBI framework to illustrate how touchscreens can offer advantages over less natural interactions. After we review relevant literature, we describe our research and present our results. We conclude with a brief discussion of the predictive utility of the elaborated RBI framework.

2 Related Work

Early empirical work with touchscreens (cf. [15]) compared them to alternative interfaces, such as mouse-based interactions, and found a speed advantage with touchscreens. More recent work has focused on the interplay between design and ergonomic features of the touchscreen (e.g. “fat finger”) on user performance. For example, these studies focus on how the features of on-screen buttons and layouts impact performance because of presumed impact on user motor interactions with the buttons and layouts (cf. [12], [13]). More relevant to the current investigation is work suggesting that user performance with touchscreens improves not simply because the ergonomic advantages they afford but because they allow for the design of closer interfaces (cf. [1], [5]). A verbal protocol study of an interior design task [11] supports the notion that closeness can improve performance. Designers using a tangible interface that allowed them to use two hands to move virtual furniture showed higher levels of spatial thinking and flexibility than those using a traditional mouse-keyboard-screen interface. The authors speculate that the more immersive environment allows the user to offload some cognitive demands of traditional interface, thereby enhancing the users’ understanding of spatial relationships involved in mental manipulation of objects. Accordingly, with tasks that involve mental manipulation of objects, touchscreens should yield high levels of performance relative to other interfaces due to both the motor control enhancement of the touchscreen *and* the reduced level of action knowledge required to accomplish some result.

3 Our Task: The Cube Comparison Task

Our task is based on the Cube Comparison Task (CCT), a context-free task that has been widely used to assess spatial ability and is believed to require mental manipulation of objects (cf. [4], [16]). In CCT, participants see two cubes with patterns on the faces; they determine whether the cubes could be the same or are different, based on the visible patterns on the sides of each cube. Patterns are not repeated. The CCT is thought to engage the components of spatial ability called visualization and mental rotation: participants visualize what is on hidden faces and mentally rotate a cube to determine if the cubes could be the same. Prior research indicates that participants of lower spatial ability, as measured on standard psychometric tests, are differentially slower and less accurate on this task as cube pairs become more complex, than persons of higher spatial ability [9]. Additionally, [8] suggests participants may differentially rotate cubes on some axes in preference to others.

4 Method and Procedure

We compared three user interaction modes with increasing levels of closeness: on-screen buttons, mouse-drag, and touchscreens. In the on-screen buttons condition, participants used the mouse to push buttons that resulted in 90° rotations. Closeness is low in this condition because the action of pushing a button is different from the actions performed in rotating an actual cube, and to push the correct button, the user must know in which of six directions the cube will turn. In the other two conditions, participants pointed to the center of a cube's face and rotated the cube by dragging the face in one of two directions using either the mouse (mouse-drag) or their finger (touchscreen). In this case, the dragging motion is similar to what might be used to rotate an actual cube. With the touchscreen, closeness is high. With the mouse-drag, closeness is only moderately high. Holding onto a mouse with a button depressed and moving the mouse on a horizontal surface to rotate a cube is not the same closeness as reaching out with one's finger and spinning the cube. In our study, participants were not required to rotate the cube and therefore could complete the task using any combination of mental and/or physical rotations using the software.

4.1 Stimuli

In our version of the CCT we placed letters on the faces of the cubes (cf. [9]). To avoid ambiguities about orientation, no letter was used that could be confused with another when rotated. The participant indicated whether the cubes could be the same or if they were different by clicking on a button marked "same" or "different."

The problems varied by complexity, defined by the number matching letters in the initial presentation and the number of 90° rotations required to make a possible match (cf. [9]). Our study had six levels of complexity. Participants completed training trials, followed by 36 trials in one of the three user interface conditions. In half of the trials, the cube pairs were the same. After the quantitative experiment, some participants were interviewed to determine their reactions to the interactions.

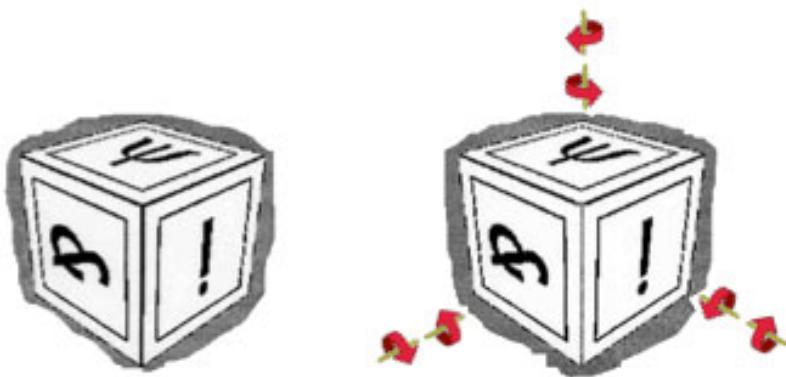


Fig. 1. The CCT with on-screen buttons

4.3 User Interaction

In our version of the CCT, users were given a pair of cubes and could manipulate the right cube via the interaction, along the x-, y-, and z-axes. We define the x-axis as the horizontal axis (off the right edge of the cube), the y-axis as the vertical axis (off the top of the cube) and the z-axis as the line-of-sight axis (off the left edge of the cube). Participants in the on-screen button condition rotated the cubes by pressing one of the six arrow-shaped buttons that depicted the direction of rotation along the three axes. The arrows were located just beyond the edges of the cubes (see Fig. 1). Participants in the touchscreen and mouse-drag conditions placed their finger or the mouse on the right cube to rotate; the “sweet spot” was on the center of the cube face (similar to Fig. 1 but without the axes or arrows).

4.4 Participants

Our sample consisted of 58 students who participated in exchange for course credit or Amazon gift cards. Thirty-one students used the on-screen button interface, 10 used the mouse-drag interface, and 17 used the touchscreen. Students also completed two standard tests of spatial ability: Paper Folding, a measure of visualization, and Cards, a measure of mental rotation [4]. Using scores that were a composite of Paper Folding and Cards, 27 students were classified as *high* and 31 *low* spatial ability. Throughout the study, we followed all required ethical participation guidelines.

5 Results

Our study had three independent variables: 1) User interface (on-screen buttons, mouse-drag, touchscreen) 2) Spatial ability (high and low), and 3) Problem type (six levels of problem complexity). We collected standard trial-level measures (e.g., accuracy, time to complete the trial) as well as those derived from the user interaction

clickstream within trials, such as time to first interaction (TFI), time between interactions, and patterns of rotation. Our hope was that the findings from the fine-grained analyzes would contribute to our understanding of the grosser measures of performance. We report first on speed-related dependent variables followed by those that are related to accuracy. Owing to differences in closeness, we predicted that touchscreen users would be most accurate, followed by users of the mouse-drag and on-screen button interfaces. From prior work [15], we predicted that touchscreen users would have faster trial completion times and shorter TFIs. Overall, we expected performance to decrease as complexity increased, both in accuracy and time, and to be poorer for users with low spatial ability.

5.1 Speed

Completion Time. Consistent with earlier work [15] we found that completion times for touchscreen users were faster than those for users of the on-screen buttons; mouse-drag users were equally fast as touchscreen users (touchscreen: 7988 ms, $SD = 2957$ ms; mouse-drag: 8026 ms, $SD = 2055$ ms; on-screen button: 9485 ms, $SD = 2082$ ms, $F(2, 52) = 3.71, p < .031$). Additionally, trial completion time differed significantly by problem type with more complex problems taking longer.¹

Time to First Interaction (TFI). TFI, the time from the beginning of the trial until the user's first interaction, is interpreted as a measure of planning time. Touchscreen users were significantly faster than the other two conditions ($F(2, 52) = 8.70, p < .005$ (touchscreen, $M = 2363, SD = 1109$ ms; mouse-drag, $M = 2581, SD = 1110$ ms; on-screen button, $M = 3661, SD = 1125$ ms). Consistent with the idea that TFI reflects planning time, TFI also differed significantly by problem type, with the more complex problems taking longer (H-F $F(34.35, 215.60) = 1.20, p = 0$).

Derived Time Per Axis. Taking all examples of the six actions (viz., two directions of rotation about three axes) and the time to execute them (i.e., the time since the previous action) allows us to compute the average *time per interaction*. This measure provides a glimpse into the relative ease of executing each action; ease of execution is a by-product of closeness and amount of action knowledge. Figure 2 shows the average time per action along the x-, y-, and z-axes (collapsed across direction) for the three interfaces. The time per interaction for the touchscreen is generally lower than that of the other interfaces. In addition, there is a significant interaction between the type of interface and the axis of rotation (H-F $F(4.0, 104.0) = 2.52, p < .045$). The interesting finding is that with the touchscreen the time per interaction is statistically the same for all three axes; the same cannot be said of the other interfaces. Executing a rotation with the touchscreen was equally easy for all three axes – just as executing rotations with a real cube would be. Whether the uneven executability profiles obtained with the other interfaces are due to motor-based idiosyncrasies or cognitive asymmetries cannot be determined, although the former seems more likely.

¹ For our repeated measures analyses of variance, we make no assumptions of the sphericity of the data and use a Huny-Feldt adjustment to the degrees of freedom, indicated by "H-F".

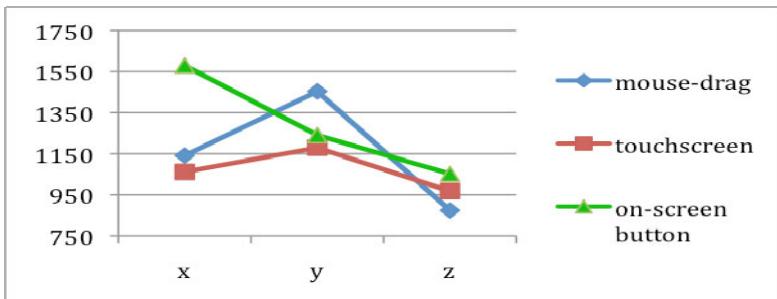


Fig. 2. Interaction of time per axis and user interface

Clustering of Interactions. From the clickstream data, we identified *clusters* of interactions based on the time intervals between interactions. For each user, we computed summary statistics for the distribution of time intervals across all trials. When an interval was longer than a user-defined threshold (viz., median interval time + (median * coefficient of variation of the distribution)) – basically, a long pause between actions -- we inferred that he or she had been on a *thinking* break, and we used that break to form the boundary of a cluster. Clusters contain actions that were executed with briefer pauses between them, and the number of actions per cluster could vary in size. The size of a cluster reflects the degree to which a user’s actions are deployed in higher-order units, suggesting that spatial and action knowledge are organized. That is, large clusters are taken to indicate bursts of action aimed to achieve multi-step goals; clusters containing single actions reflect an unplanned, step-by-step approach to doing the task. In previous research with the CCT (cf. [8]), we found that spatial ability was positively correlated with the size of clusters. In the current study, we found a similar relationship between spatial ability and cluster size, and we also found that, for our participants with high spatial ability, the proportion of larger clusters was greater with the touchscreen than with the other interfaces ($H-F(3.78, 98.28) = 2.52, p < .049$). It appears that the touchscreen facilitates the realization of spatial ability, possibly reducing the overall cognitive demands of the CCT. Perhaps the closeness of the touchscreen allows users to process longer sequences of intended rotations with fewer stoppages for thinking.

Summary of Speed Results. Overall, participants using the touchscreen were significantly faster than participants using the on-screen buttons. The speed advantage seems to spring from faster planning time and the ability to execute rotations about all three axes with equal ease. In the mouse-drag and on-screen button conditions, rotating along one axis took longer than along the other two, which would contribute to longer completion times and might cause a user to favor some axes over others and lead to less than optimal solution paths. Finally, the high spatial users in the touchscreen condition group their rotations into larger clusters, suggesting that the interface’s closeness allowed them to “see” how to do the CCT problems in larger steps by better understanding the relationships among the sides of the cubes.

5.2 Accuracy and Error Pattern Results

Accuracy overall was quite high (96.4%). There were significant differences in accuracy by type of interface ($F(2, 52) = 6.42, p < .002$) and problem type (H-F $F(2.96, 153.93) = 12.28, p = 0$), and the interface X problem type interaction was significant (H-F $F(5.92, 153.93) = 3.60, p < .002$). Somewhat surprisingly, accuracy was best with the on-screen buttons (99%, $SD = .039$), with lower accuracy in the mouse-drag and touchscreen conditions (95.2%, $SD = .041$ and 95%, $SD = .041$, respectively).

Error Patterns. Using the clickstream data, we examined sequences of rotations that appeared to be errors, such as when the cube was rotated clockwise along one axis, followed immediately by a counterclockwise rotation along the same axis, and then by a rotation along a different axis. This pattern suggested that the user initially rotated the cube about the wrong axis, reversed the first rotation to correct the error, and rotated it along a different axis; this pattern was coded as a *wrong axis* error. A rotation in one direction along one axis followed by two rotations in the opposite direction along the same axis was coded as a *wrong initial direction* error. A third pattern that we report here occurs when the user has rotated the cube in a “same” trial so that all the faces on both cubes match but continues to rotate the cube so that the cubes are no longer identical. A closer look at this *rotate beyond match* pattern shows that the extraneous rotation is nearly always in the same direction that rotated the cube into the match position in the first place: it’s akin to double-clicking in order to rotate the cube twice in the same direction.

The frequencies of occurrence of the three error patterns as a function of interface type are shown in Table 1. The occurrences of both the *wrong initial direction* and *wrong axis* patterns were lower in the mouse-drag and touchscreen than in on-screen button condition, suggesting that the users found it easier to anticipate the results of their actions with the closer interfaces. The *rotate beyond match* pattern was an order of magnitude less frequent with the touchscreen interface, suggesting that these users understood the outcome of their actions better than users of the other interfaces, knowing that they were one rotation away from match rather than two.

Table 1. Frequency of Occurrence of Error Patterns by Interface

Interface Type	Wrong Initial Direction	Wrong Axis	Rotate Beyond Match
Mouse-drag	$M = .007, SD = .020$	$M = .046, SD = .104$	$M = .040, SD = .097$
Touchscreen	$M = .007, SD = .024$	$M = .040, SD = .125$	$M = .004, SD = .016$
On-screen button	$M = .040, SD = .104$	$M = .124, SD = .222$	$M = .049, SD = .122$
	$F(2, 52) = 4.73, p < .013$	$F(2, 52) = 3.19, p < .049$	$F(2, 52) = 3.62, p < .034$

Summary of Accuracy Results. Accuracy was high in all three interface conditions, with on-screen button users being the most accurate. However, analysis of the error patterns suggests that touchscreen users had more complete action knowledge than users of the other interfaces because there was less of it to acquire. But the touchscreen users, not needing as much time to plan the next move, may have simply rushed through the task resulting in lower overall accuracy.

5.3 User Interface and Spatial Ability

An additional finding of interest, relative to spatial ability, supports the notion that close interfaces can have an impact on user cognition. Participants in the touchscreen and mouse-drag conditions took the Paper Folding and Cards spatial ability tests both before and after the CCT task. We found significant improvements of approximately 25% on standardized composite scores (-.18 pre, .13 post), following the CCT task ($F(1, 21) = 18.14, p = 0$), regardless of the participants' initial level of spatial ability. Unfortunately, we have no corresponding data for the on-screen button condition, but the results are intriguing nonetheless.

5.4 Qualitative Data

We conducted interviews of the participants in the mouse-drag and touchscreen conditions after they completed the experiment. The qualitative data indicate that participants found using the touchscreen most like actually manipulating the cubes. In describing the touchscreen, one participant said that the touchscreen "played in my head; I felt like I was really moving the cube instead of trying to get the mouse cursor to do it." Thus, the anecdotal data lend credibility to our claim that the touchscreen interface was the closer of the two user interfaces.

6 Discussion and Conclusion

Earlier work with touchscreens show them to yield faster interactions than other GUI interfaces; completion times in our touchscreen condition were faster than the on-screen button condition, but matched for speed with the mouse-drag interface. Note that the CCT is not a speeded task, with much of the time on task spent thinking rather than interacting. If speed were the only advantage of touchscreens, tasks like the CCT would not benefit greatly from touchscreen interfaces. Our findings, however, suggest that the advantage of touchscreens can be more than just speed.

Our data suggest that the touchscreen users benefitted from the closeness of the interface, as indicated by uniform ease in rotating the cube about any of the three axes and infrequent occurrence of error patterns in the clickstream. The shorter TFI with the touchscreen also suggests that touchscreen users were able to plan their actions more quickly, and the cluster size results suggest that participants with high spatial ability were able to execute their plans more efficiently with the touchscreen than with the other interfaces. The benefits of touchscreen interfaces can be cognitive in that there can be fewer cognitive demands with an appropriately designed touchscreen interface, one that allows interactions that are as close as possible to the actions performed with real objects in the environment.

Central to our concept of closeness is the idea that as interfaces become less natural, users must learn more about how to use them and what to expect from actions performed through the interface. We refer to the aspects about the interface that must be learned as action knowledge, and it's similar to what [10] calls a device model except that the devices are input-output devices instead of entire systems. As we saw with the high spatial users in the touchscreen condition, action knowledge necessarily articulates with other procedural and declarative knowledge, and it is crucial for knowing what method to select for achieving a specified goal.

Our concept of closeness borrows heavily from ecological psychology [3] in that natural interfaces are thought to afford interactions that require no cognitive supplementation. It is also consistent with ecological approaches to interface design [17] that aim to channel users' limited cognitive resources towards understanding complex systems rather than towards understanding low-level elements of the system that can be processed perceptually. Finally, we see our notion of closeness as an extension of the RBI framework and its theme of body awareness.

Of course, we need to have a better definition of closeness if the concept is to have predictive value. The three interfaces used here decrease in closeness from touchscreen button to mouse-drag to on-screen buttons partly because the physical movements used to rotate a real cube with a flick of the finger are most similar to those used with the touchscreen, followed by mouse-drag and on-screen buttons; partly because the latter two use an external device with its own operating characteristics – the mouse-- to yield rotations; and partly because the nature of the movements executed with the mouse-drag and on-screen buttons become increasingly arbitrary with respect to the pairing of an action to an outcome. The act of pressing a button is perhaps the most arbitrary; pressing a button can result in many outcomes.

This is not to say that arbitrary interfaces cannot be learned. Indeed, with little effort, a skilled typist can press a key that puts the letter S on a screen, having learned where the key is on the keyboard and that pressing a specific key yields an S rather than a D. With sufficient practice, actions can become automatic, and the unnaturalness of the interface becomes irrelevant. Yet, as the actions performed with an interface become less natural and more arbitrary, the ability to switch to a different interface becomes more of a problem. No doubt our skilled typist would have difficulty switching to a Dvorak keyboard. Echoing [2] and his call for ecological validity, we feel that as interfaces becomes less close, the actions performed become less like those performed in the natural world and more idiosyncratic to the particular device. If the goals of interface design are to permit transfer across multiple systems and ease of learning, closeness is an important variable to consider, particularly with tasks that involve the manipulation of objects.

Acknowledgments. A.J. Fuller, Jeremy Athy, Chad Parsons, Ocean Celestino, Nicole Vanduzen and Martez Mott assisted with this project. Software was written by Guy Zimmerman, Alyssa Tomlinson and Randall Littlejohn with partial funding provided by the NSF STEP Award No. 0757001, and S-STEM (DUE-0850026).

References

1. Barnes, J. G., Poor, G. M., Leventhal, L. M., Zimmerman, G., & Klopfer, D. S. (2005, September). Look and touch: The impact of touchscreens on the delivery of instructions for inherently 3D construction tasks using web-delivered virtual reality. Paper presented at the IPSI conference, Amsterdam, The Netherlands.
2. Brunswik, E. (1943) Organismic achievement and environmental probability. *Psychological Review*, 50, 255-272.
3. Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston, MA: Houghton-Mifflin.
4. Ekstrom, R.B., French, J.W., Harman, H.H., & Deman, D. (1976). *Kit of Factor Referenced Cognitive Tests*. Educational Testing Services, Princeton, NJ.

5. Hancock, M., Hilliges, O., Collins, C., Baur, D., Carpendale, S. (2009). Exploring Tangible and Direct Touch Interfaces for Manipulating 2D and 3D Information on a Digital Table. Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces, 77-84.
6. Jackson, R. L., & Fagan, E. (2000). Collaboration and learning within immersive virtual reality. Proceedings of the third international conference on Collaborative Virtual Environments, 83-92.
7. Jacob, R. J., Girouard, A., Hirschfield, L. M., Horn, M. S., Shaier, O., Soloway, E. T., & Zigelbaum, J. (2008). Reality-based interaction: A framework for post-WIMP interfaces. Proceedings of the Twenty-sixth Annual SIGCHI Conference on Human Factors in Computing Systems, 201-210.
8. Jaffee, S. D., Battaglia, D., & Klopfer, D. S. (2010, April). Tipping cubes: The effect of projection on performance and strategy on a mental rotation task in a virtual reality environment. Poster presented at the annual meeting of the Midwestern Psychological Association, Chicago, IL.
9. Just, M. A., & Carpenter, P. A. (1985). Cognitive coordinate systems: Accounts of mental rotation and individual differences in spatial ability. *Psychological Review*, 92, 137-172.
10. Kieras, D. E., & Bovair, S. (1984). The role of a mental model in learning to operate a device. *Cognitive Science*, 8, 255-273.
11. Kim, M. J., & Maher, M. L. (2008). The Impact of Tangible User Interfaces on Designers' Spatial Cognition. *Human-Computer Interaction*, 23, 101-137.
12. Lee, S., & Zhai, S. (2009). The Performance of Touch Screen Soft Buttons. Proceedings of the 27th international conference on Human factors in computing systems, 309-318.
13. Moscovich, T. (2009). Contact Area Interaction with Sliding Widgets. Proceedings of the 22nd annual ACM symposium on User interface software and technology, 13-22.
14. Poor, G. M. (2008). The effects of varying levels of reality-based interaction styles on a subject's ability to perform a 3D construction task. Unpublished doctoral dissertation, Tufts University, MA.
15. Sears, A., & Shneiderman, B. (1991). High precision touchscreens: Design strategies and comparisons with a mouse. *International Journal of Man-Machine Studies*, 34, 593-613.
16. Thurstone, L. L. (1938). Primary mental abilities. Chicago: University of Chicago Press.
17. Vicente, K. J., & Rasmussen, J. (1990). The ecology of human-machine systems: II. Mediating "direct perception" in complex work domains. *Ecological Psychology*, 2, 207-249.

TiMBA – Tangible User Interface for Model Building and Analysis

Chih-Pin Hsiao¹ and Brian R. Johnson²

¹ College of Architecture, Georgia Institute of Technology, Georgia, USA

² DMG Lab, College of Built Environments, University of Washington, Washington, USA
chsiao9@gatech.edu, brj@uw.edu

Abstract. Designers in architectural studios, both in education and practice, have worked to integrate digital and physical media ever since they began to utilize digital tools in the design process [1]. Throughout the design process there are significant benefits of working in the digital domain as well as benefits of working physically; confronting architects with a difficult choice. We believe emerging strategies for human-computer interaction such as tangible user interfaces and computer vision techniques present new possibilities for manipulating architectural designs. These technologies can help bridge between the digital and physical worlds. In this paper, we discuss some of these technologies, analyzes several current design challenges and present a prototype that illustrates ways in which a broader approach to human-computer interaction might resolve the problem. The ultimate goal of breaking down the boundary between the digital and physical design platforms is to create a unified domain of "continuous thought" for all design activities.

Keywords: Tangible User Interfaces, Computer Vision, Architectural Design.

1 Introduction

Design is a challenging cognitive activity, involving tactile, spatial, and auditory perception and “wicked problems” [2]. Buildings are almost invariably unique, inappropriate on another site, in another climate, or with other functionality and requirements. Architects must often address conflicting goals using limited information, and seeking a solution that harmonizes the parts in a process sometimes referred to as “puzzle making” [3]. The process is complex, cyclical and requires concentration, making it prone to disruption from external sources as well as internal ones.

During the process, architects use external media to record and explore their designs. Traditionally, pen and paper were used for drawings, and cardboard, wood, clay and other materials were used for models. While most work is now done on computers using CAD and modeling software, architects in the early stages of design are often more comfortable directly creating or modifying hand drawings and concept models than using software. A model may be used directly to study physical issues such as shadowing and lighting; it can be used as the focus of face-to-face conversations with clients who might have difficulty interpreting architectural drawings and digital models, and the designer can trim or replace parts in order to

study alternatives. This does not mean they are unaware of the value of digital tools. Digital models are known to be useful for high quality visualization, extraction of coordinated drawings and cost estimates, and key to energy and lighting simulations, etc. Unfortunately, the issues are often quite subtle, the typical “direct manipulation” interface [4] provides limited freedom for designers to study their designs, and it remains difficult to edit in 3D on a 2D screen [5].

The apparent need to choose between two high-value tool sets motivates this vision: your computer ‘partner’ watches over your shoulder as you construct a cardboard model and simultaneously builds a digital model. You would have the physical artifact and the digital data at no extra cost. In fact, the two might be used together to provide even greater insight into the design as it is being developed.

2 Related Work

In 1997 Hiroshi Ishii proposed Tangible User Interfaces (TUIs) as a new way to bridge between “Bits and Atoms” [6]. He argued that information “bits” would come into the real physical world through everyday physical objects and environments (atoms) [6]. By making digital information tangible, he thought the gap between cyberspace and physical environments could be bridged. During design, spatial cognition plays a very important role, helping us understand complex relationships and manipulate abstract concepts. Interruptions can disrupt continuity of thought and complex GUIs frequently present command-selection interruptions. However, Tangible User Interfaces may enable us to better apply our intelligence during the design process.

There have been several previous efforts to create tangible interfaces for design processes. In “URP”, Underkoffler and Ishii address the issue in the context of urban planning by combining tangible models and projected simulation data [7].

There is other research that addresses the relationship between the physical and digital geometries by using smart materials. Construction kits, such as LEGOTM Technic, are easy to use [8]. Anderson *et al.* made a project that combines building blocks and a scanning system to achieve the goal of modeling and reducing the inaccessibility of the traditional sketch modeling system. “FlexM” is another project that uses two kinds of construction kits to build and shape a 3D digital model [9]. Researchers developing “Posey” utilize embedded sensors to make construction kits more complicated and adaptable to design [10]. Instead of using smart materials, Song *et al.* use digital pen for recording annotation and editing geometry in both physical and digital world simultaneously [11]. However, in these systems, the digital geometries have to be pre-built and registered to the particular physical objects, which is not desirable for designers in the midst of the design process.

3 The TiMBA Prototype

3.1 System Design Considerations

Cardboard models are commonly used by architects in studying the form of a building. To construct a digital model from a physical one, our “over the shoulder”

software system has to know the shape, the location and the orientation of the individual pieces of cardboard. There are several possible ways to find the whereabouts of physical objects for the system, such as embedding radio frequency ID tags in every piece of cardboard, building a model with multiple 6 degree-of-freedom sensors, or detecting the related position of every piece by knowing which are attached, such as the improving version of Posey [10]. However, these all restrict the range of formal expression available to the designer. In the end, we adopted a computer vision strategy in order to work with almost any cardboard, permit reshaping of pieces, and minimize instrumentation costs (as compared to processing time) while producing a practical system. We divided the model-building process into two steps. The first step employs an edge detection algorithm to define the shape of each cardboard piece. The second step uses the marker-tracking library from ARToolKit [12] to locate each cardboard piece in the physical assembly.



Fig. 1. Foundation board for users to build model on

Since cardboard pieces are usually cut flat in the first place, we provide a black cutting mat with an overhead camera for a better edge detection process. After cutting, each piece is given a fiducial marker. The camera acquires an orthogonal, or “true-view” of the piece and links it to the marker ID and location on the piece. After users combine the individual cardboard pieces in step two, the composite digital model can be computed using the positions of the markers relative to a marker set on the foundation board (Figure 1). As the designer works, a digital model is constructed from their physical model. If the designer decides to reshape a piece, they only need to remove it from the model and adjust the size on the cutting mat before re-inserting it in the composition.

3.2 System Overview

In this prototype, we modified ARToolKit 2.7 (ARTK) to address both vision tasks, and used an API for the commercial modeling program Google SketchUp to build the host-modeling environment. ARTK is designed for use in Augmented Reality applications. TiMBA utilizes the ARTK edge-detection code and the ability to register the location and orientation of fiducial markers. Two applications based on

ARTK were created to complete the two steps identified above. Step 1 is controlled by the “Shape Scanning Application” (Figure 3) and Step 2 is controlled by the “Location Detector Application” (Figure 4). Each application runs continuously, receiving frames from its respective camera and performing its analysis. They transfer results to a custom script running in SketchUp, which combines the results. Our SketchUp script, “SketchUp Ruby Helper” (SRH), receives the data provided from the two ARTK applications and uses it to control the shape, location, and orientation of the geometry of the finished digital model. Figure 2 shows the whole system architecture.

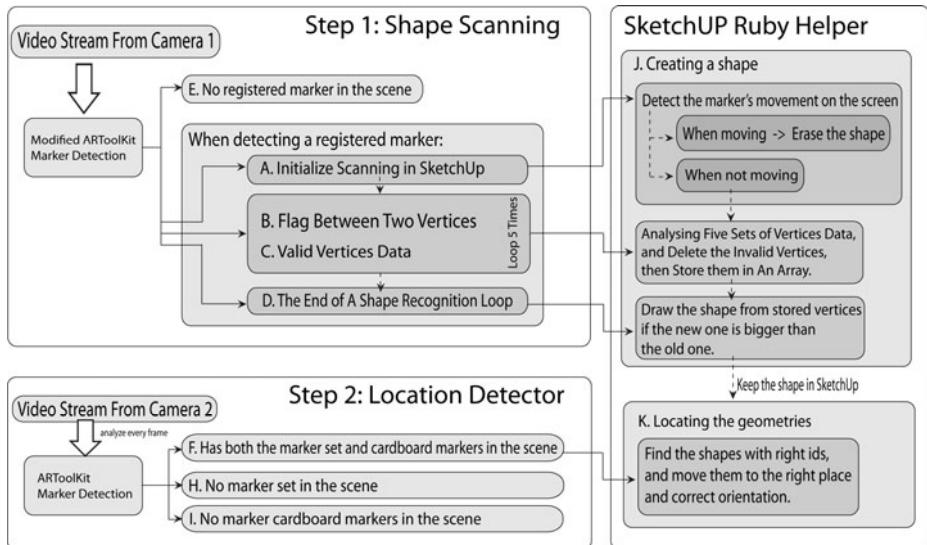


Fig. 2. The System Architecture of TiMBA

3.3 Step 1 - The Shape Scanning Application (SSA)

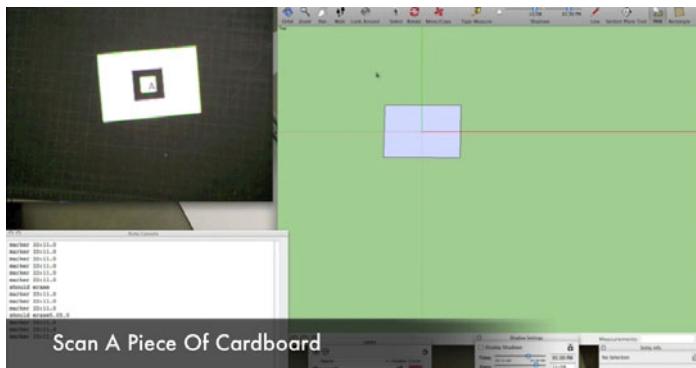
The Shape Scanning Application (SSA) uses a utility process from ARTK to accumulate the dark area, and gives the same area a flag. In the original ARTK process, these are recognized whether they are a four-vertex shape or not. In our modified version, the new code grabs the area information of reversed color to recognize the number of vertices around the contour of the area and which pixels these vertices are when the previous frame contains a fiducial marker. In order to get more accurate vertex information, TiMBA combines information from a number of video frames.

In addition to the vertex information, SRH needs to know the size of the shape, the ID of the marker, and the transformation matrix of the marker. These data, combined with the vertex data will comprise an entry in the complete data stream. A Unix “pipe” is used to communicate between SSA and SRH. Each cardboard piece is represented as a record similar to the following:

Table 1. Sample record of data transferred from SSA to SRH

-1	2	383.7	340.1	170.2	277.7	164.5	197.9	268.8	187.6	283.2	264.1
a	b	c	d	e	f	g	h	i	j	k	l

In Table 1, the meanings of these numbers are as follows: (a) beginning of data flag, (b) the marker id, (c) marker center X, (d) marker center Y, (e) to (l) the coordinates of the four vertices of the marker on the XY plane of the piece of cardboard.

**Fig. 3.** The Screen Shot in Step 1

3.4 Step 2 - The Location Detector Application (LDA)

The concept of this application is simply to compare the location data of the cardboard markers to the coordinate system provided by a “base” marker set on the foundation board (Figure 1). Results are transferred to SRH to drive the model display. To improve the speed when multiple markers are encountered in the same scene, a continuity assumption is made: in the absence of contraindication, parts do not move. For each frame analyzed, the system compares the location data of the markers it sees to that of the previous frame. If a marker has not moved too much, it will not send new location information for the marker, but send the next moving marker in the current loop instead.

3.5 The SketchUp Ruby Helper (SRH)

This program has to analyze the data from both LDA and SSA. Then, it transforms them into the coordinate system of SketchUp and displays the geometry for users. There are three main tasks to deal with in this application: 1) creating a 2D shape, 2) moving the 2D geometry to the appropriate 3D location, and 3) trying to glue the adjacent shapes together. The first two tasks simply take the data from the other two applications and analyze them to verify a reasonable shape and location. The goal of the third task is to assert physical continuity in the system—to assemble the cardboard

pieces as one big model, seeking to avoid the situation where, because the camera for the LDA cannot see a marker at the rear of the model but the user's viewpoint shows it, pieces of the digital model "drop out".

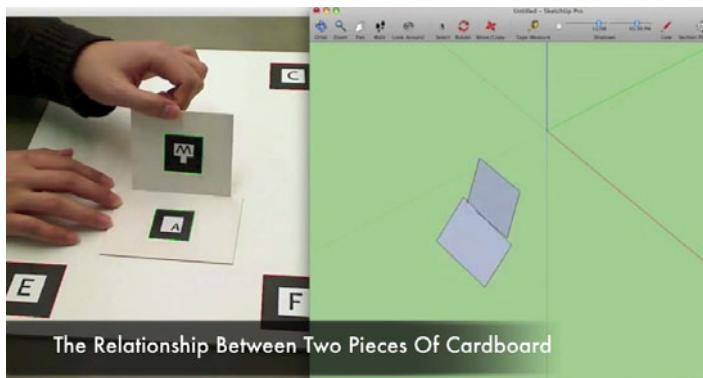


Fig. 4. The Screen Shot in Step 2

3.6 Discussion

This version of TiMBA helps the user make a digital model by "observing" their construction of a physical model. It can reflect the movement, shape, and position of a piece of cardboard in the physical world and give some response back to the user. However, the real benefits of this kind of system reside in the synergy that might be achieved between the physical and the digital by utilizing the ability to give much more information and have a better way to represent the information than the physical model can provide. Before this happens, we believe that there are some improvements we need to make. Even though the prototype provides real time information from the physical side, response-time is inadequate. In SSA, after the camera sees a marker, it takes a while to "grow up" the right shape in SketchUp. Likewise, in LDA, the movement of the digital geometry, driven by movement of the physical cardboard, is not smooth enough. Accuracy is another issue. Since the system has to run in real time, the result of the computer vision analysis is not always as accurate as in a static scene. Although TiMBA includes efforts to mitigate this error by "gluing" or linking nearby edges from two different shapes, this issue would still be an irritant to users. Self-occlusion is another issue alluded to previously. The physical model is opaque. As pieces are added, they may obscure the camera's view of other pieces. Currently we rotate the model base to enable the camera to see all of the model. An additional camera view of the model would enable the system to run more smoothly, but self-occlusion remains an issue. While imperfect, we believe that this prototype is just the beginning, and there are still many more possibilities we could explore.

4 Preliminary Experiment

Observations of users are always valuable because everyone has his/her own habits when building models. Therefore, after we built the prototype, we asked several

students from our architecture school to test it. While TiMBA uses a general model-building approach intended to fit every designer's habits, we had to add AR marker placement and scanning to the workflow. Further, we wished to test the utility of the system to help users answer certain architectural questions through modeling.

At the beginning of the tests, the participants were given a short set of instructions and permitted a 5 minute "learning phase" so that they could familiarize themselves with the system. We provided the same tools and materials for each participant to complete both a digital and a physical model using TiMBA. After finishing the learning phase, they were given two tasks, including the building of digital and physical models by provided drawings and the answers to two related questions. For these tasks, the subjects were asked to consider a particular design of an element for this building. We also avoided discrete, binary, yes/no style answers for tasks, which might encourage guessing. Subjects were told they were being timed on the experiment and to "complete each task as quickly as possible." After building the models, they were asked to save the digital model and kept the physical one for the experiment result, and then moved on to the next experimental condition. At the end, they were given a post-experiment questionnaire.

We employed both performance measures and subjective perceptual measures. In performance measures, we recorded a) time (the number of minutes to complete the test) and b) the accuracy of the results. Assessing accuracy, we compared their answers to the task questions with a correct answer. For example, the first task is "at 12:00 pm, June 21, which walls would receive direct sunshine (east, west, south, north)?" The subject had to build an accurate model and then use the digital model to simulate the shadow patterns on the site. An alternative question was "which orientation of the building is the best one in terms of avoiding the shadows from the other building near the site for the garden near the building (east, west, south, north)?" In this case, the subject could rotate the digital model until they achieved a satisfactory answer. The results are convincing. All of our subjects achieved the correct answers in a relatively short time when constructing both of the digital and physical models.

Figure 5 shows the list of questions and the results from the post-experiment questions. In general, the subjects did not think "sticking the marker" on the pieces distracted them from building the physical model. They also believed that the assistance from the computer helped them answer questions about the design. However, there are still some features that bother users: the prototype requires users to scan each piece of cardboard before assembling them, a process which takes several seconds. This annoyed subjects who are used to gluing two pieces of cardboard together right after cutting them. In addition, users must pay attention to which side of the cardboard they affix the AR marker to, because the system needs to see the marker to locate the geometry. They also think the prototype could be more intuitive. According to the observation, the separation between the working and data representation platforms could cause the intuitive issue with which the subjects encountered. Although they thought they could build a better digital model with mouse and keyboard, some of them indicated that it depends on the complexity of model. For instance, some complicated models require advanced digital modeling skill. On the

physical side, however, we only need our hands to bend materials or place them at the specific location. Overall, they think this is a fun and effective way to build a digital model.

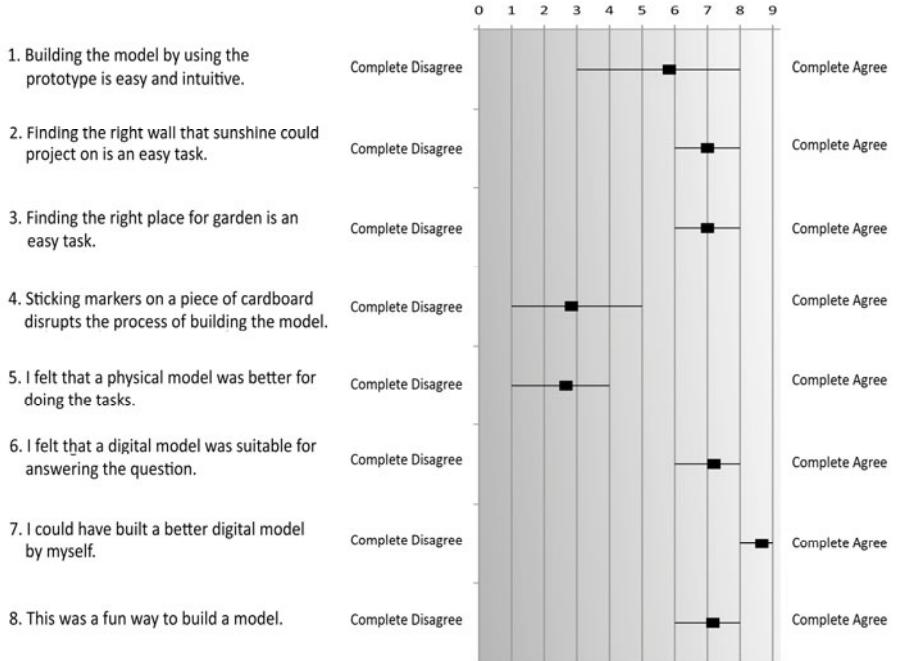


Fig. 5. List of the Assessment Questions and the Results

5 Conclusion

In this paper we have presented some recent uses of Tangible User Interfaces and other digital technologies, how they are increasingly important in the design process, and how, by using these new technologies, we may derive certain benefits. We also explained why we believe designers will not totally abandon the traditional design media, which rely on hand drawings and physical models. Finally, we described TiMBA, a prototype software system that we developed to demonstrate that the affordances of both physical and digital workflows might be combined synergistically. The most exciting “take-away” from these applications is the notion that we can treat a physical object in digital ways. This could lead to infinite possibilities when computational power can get “off the screen” and lay on our working platform.

Why do we not simply finish all the tasks in a virtual world? Why do we need to connect the physical and virtual worlds together? We believe that the ubiquity of computational power will let us treat computational tools as additional physical objects. These two different categories of “objects” are essentially the same thing,

created and utilized by us. We seek to make available the affordances of every tool designers wish to utilize to achieve their goals. The computational power is just the magic that can make these objects alive and offer them the ability to interact with humans, enabling them to become advisors when we need them. Ambient displays and tangible interfaces suggest the road forward.

All of these new approaches make the design process less cumbersome. At the present time, users struggle to manipulate digital objects though they know how to interact smoothly with physical ones. With the widespread use of TUIs, the boundary begins to blur between digital and physical worlds. When any physical object can be instantly given a digital representation and any digital data may be made evident in the physical world, the two begin to merge. Using computer vision, all objects in the physical world, not just the ones with sensors, can become part of our digital environment. The designer, able to work in either or both, experiences “continuous thought” and produces better design.

Acknowledgements. We thank Daniel Belcher, Randolph Fritz, members of DMG Lab at University of Washington and the colleagues at Georgia Tech for giving us insightful feedback and suggestions.

References

1. Binder, T., Michelis, G.D., Gervautz, M., Jacucci, G., Matkovic, K., Psik, T., Wagner, I.: Supporting configurability in a mixed-media environment for design students. *Personal Ubiquitous Comput.* 8(5), 310–325 (2004)
2. Horst, R., Webber, M.: Dilemmas in a General Theory of Planning. In: *Policy Sciences*, vol. 4, pp. 155–169. Elsevier Scientific, Amsterdam (1973)
3. Archea, J.: Puzzle-making: what architects do when no one is looking. In: Kalay, Y.E. (ed.) *Principles of Computer-Aided Design: Computability of Design*, pp. 37–52. Wiley-Interscience, Hoboken (1987)
4. Shneiderman, B.: Direct manipulation: a step beyond programming languages. *IEEE Computer* 16(8), 57–69 (1983)
5. Hall, R.: Supporting Complexity and Conceptual Design in Modeling Tools, in *State of the Art in Computer Graphics: Visualisation and Modeling*. In: Rogers, D.F., Earnshaw, R.A. (eds.) 153 – Quotation from R. Sproull, Keynote Speech, SIGGRAPH 1990. Springer, Heidelberg (1991)
6. Ishii, H., Ullmer, B.: Tangible bits: towards seamless interfaces between people, bits and atoms. In: *Proc. SIGCHI 1997*. ACM Press, New York (1997)
7. Underkoffler, J., Ishii, H.: URP: a luminous-tangible workbench for urban planning and design. In: *Proc. SIGCHI 1999*, ACM Press, New York (1999)
8. Anderson, D., Frankel, J.L., Marks, J., Agarwala, A., Beardsley, P., Hodgins, J., Leigh, D., Ryall, K., Sullivan, E., Yedidia, J.S.: Tangible interaction + graphical interpretation: a new approach to 3D modeling. In: *Proc. 27th Annual Conference on Computer Graphics and Interactive Techniques*. ACM Press/Addison-Wesley Publishing Co., New York (2000)
9. Eng, M., Camarata, K., Do, E.Y.-L., Gross, M.D.: FlexM: Designing A Physical Construction Kit for 3d Modeling. In: Akin, O., Krishnamurti, R., Lam, K.P. (eds.) *Generative CAD Systems: Proc. GCAD 2004 International Symposium on Generative CAD Systems*, Carnegie Mellon University (2004)

10. Weller, M.P., Do, E.Y.-L., Gross, M.D.: Posey: instrumenting a poseable hub and strut construction toy. In: Proc. 2nd International Conference on Tangible and Embedded Interaction. ACM Press, New York (2008)
11. Song, H., Guimbretière, F., Hu, C., Lipson, H.: ModelCraft: capturing freehand annotations and edits on physical 3D models. In: Proceedings of the 19th Annual ACM Symposium on User Interface Software and Technology, Montreux, Switzerland, pp. 13–22. ACM, New York (2006)
12. HIT Lab. (n.d.). ARToolKit, <http://www.hitl.washington.edu/artoolkit/> (retrieved March 1, 2009)

Musical Skin: A Dynamic Interface for Musical Performance

Heng Jiang¹, Teng-Wen Chang², and Cha-Lin Liu³

^{1,2} Graduate School of Computational Design, National Yunlin

University of Science and Technology

³ Department of Multimedia and Animation Art, National Taiwan University of Arts
`{g9834706, tengwen}@yuntech.edu.tw, liuchalin@gmail.com`

Abstract. Compared to pop music, the audience of classical music has decreased dramatically. Reasons might be the way of communication between classic music and its audience that depends on vocal expression such as timbre, rhythm and melody in the performance. The fine details of classic music as well its implied emotion among the notes become implicit to the audience. Thus, we apply a new media called dynamic skin for building up the interface between performers and audiences. Such interface is called “Musical Skin” is implemented with dynamic skin design process with the results from gesture analysis of performer/audience. Two skins-system of Musical Skin are implemented with virtual visualization/actuators/sensible spaces. The implementation is tested using scenario and interviews.

Keywords: Dynamic skin, Sensing technology, Musical performance, Scenario.

1 Introduction

In human computer interaction, we often focus on how computer amplifies human activities. Among those activities, music performance has smaller percentages due to its complex and emotional expression. Until now, only a few conferences or researches aim at exploring the musical expression with computers, namely NIME. However, musical expression did provide a strong interface that can touch and move the audiences’ heart. With the state-of-the-art technology in HCI, how to address this issue in terms of human computer interaction shall provide an effective interface for bridging the performer and audience.

1.1 Musical Expression as an Interface between Performer and Audience

Compared to pop music, the audience of classical music has decreased dramatically. Reasons might be the way of communication between classic music and its audience that depends on vocal expression such as timbre, rhythm and melody in the performance. The fine details of classic music as well its implied emotion among the notes become implicit to the audience. Thus, a good performer is the key to the music performance that cannot only perform the music according to the notes but also the emotion among them. With the far distance of stage, audience can only access the

music based on its result not the process. Therefore, a dynamic interface for sensing the inner feeling of performer and corresponding to the reaction of audience is an interesting and valid HCI research.

1.2 New Media Interface: Dynamic Skin

For implementing a dynamic interface between performer and audience, we adapt the concept of dynamic skin from architectural research. Dynamic skin starts with a façade system that can react according to the environment surrounding the building. As the figure below (figure 1) is a dynamic tensile structure, a unit can be constituted by this movable tensile structure; while adding a microcontroller it can be controlled by programming to perform corresponding variations [1]. The responsible facade is obtained more and more attentions; the inventions of the sensor and the actuator just fulfilled the requirements for designers[2].

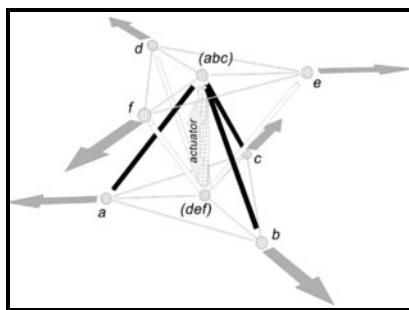


Fig. 1. Dynamic structure

For example, FLARE (2009) is an architectural dynamic skin unit system developed by WHITHvoid in Berlin, Germany. It can clad the architecture and surface of wall, and control one unit by one metal sheet and cylinder. The computer manages system and sensors to communicate with outside and inside erections through ambient light, and lengthen and shorten FLARE unit by cylinder to arise visual effects as figure 2 [3].

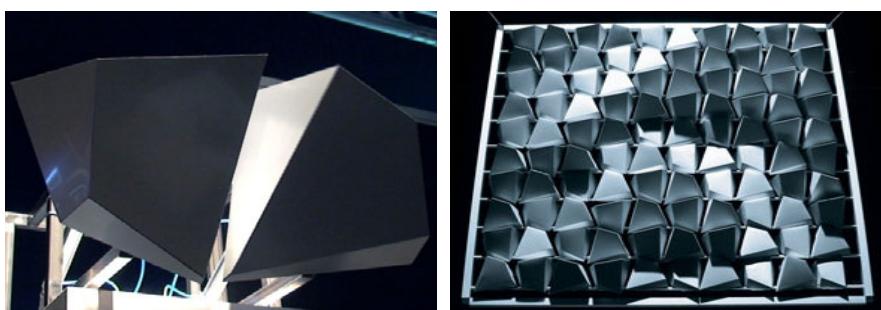


Fig. 2. FLARE-façade

With exploring the interface part of skin, Hyposurface provides another dimension to look at the performance of dynamic skin. Argis Hyposurface is a kind of responsive architectural dynamic skin unit driven angle and position of thin metal sheet on appearance to by individual mechanical structure. It convey the a type of wavy experience and feedback of dynamic skin through variation of position and form of whole thin metal sheet to create visual effect on curved surface as figure 3 [4].

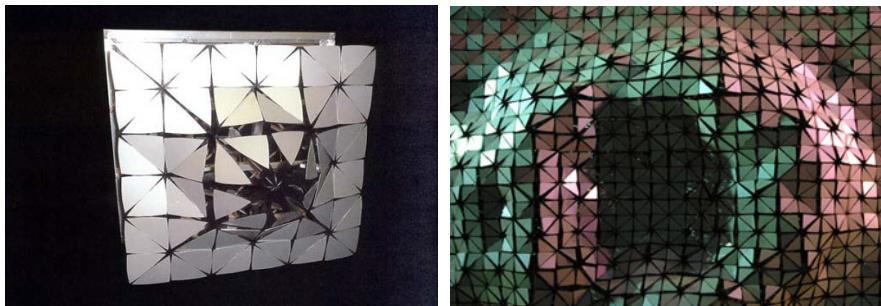


Fig. 3. Aegis Hyposurface

This structure of Hyposurface utilizes units to constitute the construction of wall and build dynamic skin with large area. After receiving signals, dynamic skin can do diverse change from point to face. The conceptual analysis of this case has enable structure of dynamic skin generating displays pattern or outline to attain wavy effect.

1.3 An Augmented Interface Based on Dynamic Skin Is Needed for Supporting the Communication between Performer and Audience

With the dynamic skin technology including sensors and actuators, a new type of interface will not be limited by its physical form but be capable to sense the ability of performer and reactive feedbacks from audiences. The interaction technique of this dynamic interface is based on the input and output of the computational unit, namely it is accomplished by the signal transmitting mechanism between human and interface [5], illustrated by figure 4 below:

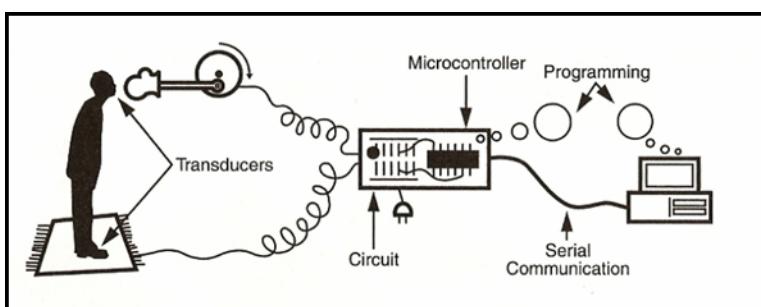


Fig. 4. The mechanism of signal transmission

In this research, we are aiming at exploring an augmented interface based on the dynamic skin technology to support the communication between performer and audiences.

2 The Problem

For exploring the computational feasibility of an augmented interface using dynamic skin technology, the problem of this research is divided into two folds: how to convert sound-based expression (music) to physical expression (dynamic skin)? And how to ensure such two-way communication between performer and audience from the skin developed?

By adapting the basic technology of dynamic skin: sensible structure and interaction design process, the approaches conducted in this research are (1) sensor-based approach for gathering the information from both performer and audience, (2) using interaction system design process to model the interactive communication between performer and audience.

The first approach: augmented music expression in dynamic skin, sensor technology is used to collect possible data from both performer and audience during the performance. Hence the analysis and encoding over collected data, and then transform into visual performance shown on two skins—one is onstage, and another is façade.

The Second approach: interaction design for modeling the interaction between performer and audience along with music itself. Using interaction design process, the each interactive process (music-skin-audience, performer-skin-audience, music-skin-performer) is designed and analyzed.

3 Analysis

Performers and audience' physical changes will reflect onto their behaviors, such as skin temperature, heart beats, voice, body movements and so on. Since these onstage or offstage behavioral changes can be measured and analyzed, we can utilize this body information to interact with the skin.

The responsive technology can be divided into sensing and actuators. The sensing technology transforms the space into a sensible space. Such spaces can then covey physical information such as chemical, biomass and so on into the possibility of re-usable electronic devices and output signals. Sensing technology blurs the relationship between the physical and virtual information, and led us into an irregular, but the rich concept of innovative interface [6].

With approaches above, the correct mapping between behaviors and perform-related information is needed. Thus, we need to gather correct information mapping in three catalogues: performer, audience and dynamic skin structure design.

3.1 Performer-Based

In this stage, we need to collect the mapping between music to emotional expression as well as the subtle changing variable of performers physically (due to the criteria sensor technology) while on stage. Interviews and behavior observation methods are

conducted for understanding what kind of sensors can be used to collect desired information. Questionnaires over emotional expression are also conducted for analyzing the possible mapping.

3.2 Audience-Based

On the audience side, how audience behaviors and reacts is the targeted mapping for study. Since classical music is often conducted in concert hall, the audience outside the concert hall will be the challenge for this stage. Two steps: the audience behaviors during the performance in the concert hall, and the reaction after the performance are observed and recorded. Additionally, unofficial interviews are conducted for exploring the mapping and further developing the interaction process for this research.

3.3 Dynamic Skin Structure

With the mapping and interaction process analyzed from previous stages, the structure of dynamic can then be designed and simulated before implementation. Three Steps: first, dynamic structure adjustment for containing the components developed in second step. Second step is to transform the mappings in designing the required sensors and connection wires for triggering the interaction and gathering the information from either performer or audience. Third step is to explore the connection with network due to the two-skin designed, The Mapping Mechanism as the table 1.

Table 1. Mapping used in the prototype: musical skin

Sensing unit	Sensor	Item	Computing unit	Signal	Feedback
Music	Microphone	Music	Sonic processing unit	Volume	On onstage and on façade: unit vertical rotation
				Pitch	On onstage and on façade : unit horizontal rotation
Performers	Polar Wearlink®31 heartbeat sensor	Heartbeat	Polar heartbeat Receiver HRMI	ECG	On onstage and on façade : Light color (Red↔White)
	Temperature sensor	Temperature	processing unit	Digital signal	On onstage and on façade: Light Brightness (Bright↔Dark)
	Optical sensor	Movement	processing unit	Digital signal	On onstage and on façade: Skin wave
Audience	Voice sensor	Voice	processing unit	Analog signal	On onstage: Flash light On façade: Ambient light

4 The Solution - Musical Skin

With the analysis from performer, audience and dynamic skin structure, we implement a prototype called “Musical Skins” that have double skins system: one is on stage and another is outside the concert hall, namely onstage skin and façade skin. Onstage skin collects the information from performer and music, and then outputs the feedback from audience. On the other hand, façade skin will collect the information from audiences and output the performance via physical skin and the music speaker. The system diagram of Musical Skins is shown in Figure 5.

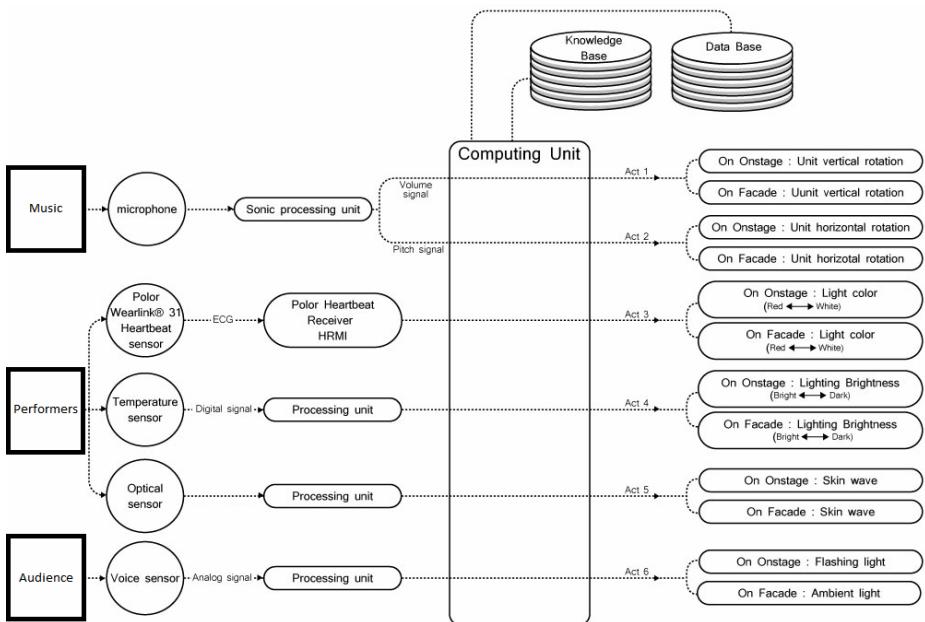


Fig. 5. System diagram of Musical Skin

4.1 System Design

Musical Skin is comprised of three parts (a) sensors collecting the signals, (b) computing unit processing collected signals and (c) output to dynamic skin for performance. Different input will generate different response from skin. The whole system diagram of signal/information flow is shown in figure 5.

The exemplary signal process is shown in two signals: heartbeat and sound sensor as followed input/computing/output.

1. Input: heartbeat and sound (volume and rhythm of music) signals from performer and music are used as the trigger for the system. Once signals received, system will start to collect the sound signals from music until it is done; for heartbeat, performer

will carry wireless heartbeat sensor and system will receive the first 5 min of calm condition as initial data, and based on this initial data to distinguish the variation of heartbeat later on during the performance.

2. Computing: while sound signal received, computing unit will use Fast Fourier Transform (FFT) to analyze/encode the signal, and send them to the controller (arduino) of dynamic skin for actuate the performance of skin. Similarly, heartbeat signal from performer also went through analyzer and encoder before sending the skin to actuate the color change of RGB LEDs. (3) Output: Once signals are encoded and sent to the actuators, the different skin units will response to different signals received. The movement as well as lights is all part of skin units that will have act and perform just like the skin is the collection and reflection of all signals received from performer/audience/music. Onstage skin will react and give feedbacks to performer and play as part of performance, and façade skin will incorporate with the speaker/music as part of visualization/musical expression of performance that will give feedbacks as well collecting information from audiences.

4.2 Sensor Components

Sensor components are classified in three groups: performer, music and audience. Different group will have different technology based on their activities. (1) Performer: the sensors for performers are based on their physical information while performing on stage; they are (a) heartbeat sensor, (b) temperature sensor and (c) optical sensor. Each sensor has its own wireless feature such that it will not disturb the performance. (2) Music: the sensor for music is sonic sensor, using microphone to collect the sound and computing unit as synthesizer to analyze/encode the signals. (3) Audience: voice sensor is used for sensing the volume of audience's response and optical sensor for gesture of audience.

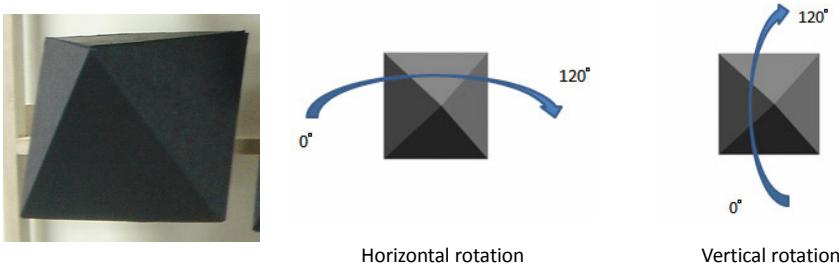


Fig. 6. Unit and unit movement

4.3 Motion Components

The dynamic skin is comprised of a set of motion components (as shown in figure 6). Each unit is a pyramid with width 17.5 cm, length 17.5 cm and height 14 cm. 4 x 4 units form a cluster for dynamic interface (as shown in figure 7).

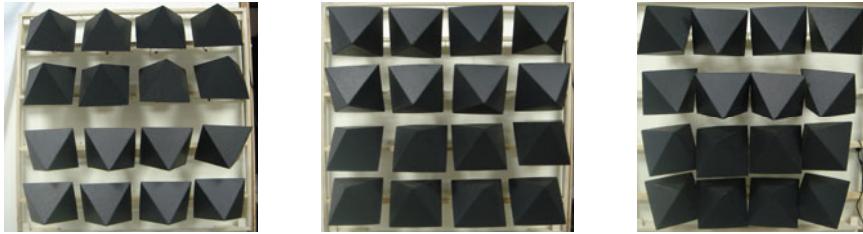


Fig. 7. Cluster for dynamic interface

The cluster receives the signal from computing unit and mapping to the different performance of skin according to the rule-bases. The power source of dynamic skin is servo and each unit is controlled by two servos located at the back of unit. The controller used in this prototype is arduino that controls two servos for vertical and horizontal rotation of units (as shown in figure 8).



Fig. 8. Double servo motor control status of horizontal and vertical rotation

5 Scenario

When we hear music playing, system captures the heartbeat, temperature and posture of audience and sense the playing music. Skin begins to perform this moment and units swing in up, down, right and left four directions by size of volume and variation of tune. It reinforces our feeling for admiring music tempo and strain. System sense heartbeat of performer turns into fast by spirited part of music, light on skin will convert from white into red to present performer's emotion soaring. Skin system also sense the body moving of performer to wave and the red light become brighter to show the summit of emotion of music.

After performance finished, audience are satisfied with and get up to applaud. System sense the sound of applauding and there glitter light on wall of hall. While performer sees the twinkle to realize this show is successful and encouraged by everyone to be willing to give more shows. People outside of performance hall will stop to appreciate due to the wonderful performance of musical skin.

6 Conclusion

This system discovers structure make skin interface act not obviously when testing and evaluation on the aspect of structure design. There some delays on serve as

derivation of performance skin. This is for the sake of serve belongs to mechanical device which can't communicate with skin directly, so the structure of unit must be improved.

6.1 Findings and Contribution

We found inaction problem in musical performance from observation of music admiration. To solve problem, the method is to design a dynamic skin correspond the music playing scenario. On observation and interview both sides arrange adaptive sensing technology on skin to make interface sensing and giving feedback. This all is for the target to do three communications of performance and audience, performer and dynamic skin, audience and dynamic skin so add these three communications into to create new user experience.

We realized how take the process of design of dynamic skin as foundation in this study. There is a further promotion in design of structure and movement of skin via inference of musical skin to bring up a kind of design process of musical skin.

6.2 Future Works

As current musical skin system, the interface still stays in testing stage and perhaps can't correspond to emotion performer want to convey to audience. Consequently, we will keep going to amend unit of sensing and structure/appearance of interface to do better communication between performer and audience.

References

1. Sterk, T.: Using Actuated Tensegrity Structures to Produce a Responsive Architecture. In: ACADIA, pp. 85–93 (2003)
2. Hu, C., Fox, M.: Starting From The Micro A Pedagogical Approach to Designing Responsive Architecture (2003)
3. WHITEvoid interactive art & design, <http://www.flare-facade.com/>
4. Goulthorpe, M.: AEGIS HYPOSURFACE©. Routledge, New York (2001)
5. O'Sullivan, D., Igoe, T.: Physical Computing: Sensing and Controlling the Physical World with Computers. Thomson Course Technology (2004)
6. Tsai, C.-C., Liu, C.-L.C., Chang, T.-W.: An Interactive Responsive Skin for Music Performers, AIDA. In: Beilharz, K., Johnston, A., Ferguson, S., Chen, A.Y.-C. (eds.) Proceedings of the 2010 Conference on New Interfaces for Musical Expression (NIME 2010), pp. 399–402. The University of Technology Sydney, Sydney (2010)

Analyzing User Behavior within a Haptic System

Steve Johnson¹, Yueqing Li¹, Chang Soo Nam¹, and Takehiko Yamaguchi²

¹ University of Arkansas, Fayetteville, AR, United States
`{sajohns,yx1002,cnam}@uark.edu`

² Laboratoire d'Ingénierie des Systèmes Automatisés (LISA)
Université d'Angers, 62 Avenue ND du Lac – 49000 Angers, France
`{takehiko.yamaguchi}@univ-angers.fr`

Abstract. Haptic technology has the potential to enhance education, especially for those with severe visual impairments (those that are blind or who have low vision), by presenting abstract concepts through the sense of touch. Despite the advances in haptic research, little research has been conducted in the area of haptic user behavior toward the establishment of haptic interface development and design conventions. To advance haptic research closer to this goal, this study examines haptic user behavior data collected from 9 participants utilizing a haptic learning system, the Heat Temperature Module. ANOVA results showed that differences in the amount of haptic feedback result in significant differences in user behavior, indicating that higher levels of haptic friction feedback result in higher user interaction proportions of data. Results also suggested that minimal thresholds of friction haptic feedback can be established for a desired level of minimum user interaction data proportions, however; more research is needed to establish such thresholds.

Keywords: Haptic User Interface, Thermal Device, User Behavior, Inclusive Design.

1 Introduction

Haptic technology enables users with the ability to touch, feel, and interact with virtual objects in a computer simulated virtual environment as if they were real, physical objects. Haptic devices, coupled with sophisticated software applications, allow the user to interact with haptically rendered objects exhibiting a variety of surface textures and hardness in both two and three dimensions [2]. Haptic systems consider the unique physiological and psychological attributes related to the sense of touch and utilize haptic feedback to render haptic sensations [12]. Such haptic feedback is achieved through the haptic device via tactile, force, and torque feedback [6]. Haptic feedback tools such as gravity wells, indents, dimples, ridges, friction, and dampeners have emerged and are generally accepted among the haptic community with the specific purpose of providing users with anticipation, follow-through, indication, and guidance [14].

Haptic technology has the potential to enhance education, especially for those with severe visual impairments (those that are blind or who have low vision), by presenting

abstract concepts through the sense of touch. As an assistive learning tool, haptic technology has proven to be effective - enriching and improving students' learning and problem solving ability by evoking haptic stimuli and sensorial feedback [3; 5; 11; 23], combining audio and haptic perceptions [7; 13; 25], and supporting tangible investigation and experimentation with concrete objects [10; 16; 18]. Likewise, the implementation of haptic force feedback has been shown to influence task performance and sense of copresence [1; 8; 21]. Finally, the natural pairing of haptic technology with the prevalence of virtual education research provides further building blocks towards combining learning with leisure [19].

Despite the advances in haptic research, technology, human sensing and control [4; 9], and haptic behavior [20; 22; 24], little research has been conducted in the area of haptic user behavior toward the establishment of haptic interface development and design conventions [5; 6]. This lack of research can be largely attributed to the cost, multi-domain knowledge, time, and effort associated with the design and development of haptic systems [15]. By proposing an acceptable method of analyzing haptic user behavior, it may be possible to establish proper methods in which to evaluate haptic technology, haptic methods, and haptic interface design and development. The overall goal of such a methodology is in maximizing ease of use in the domains of haptic interaction and haptic user interface design for both sighted and visually impaired users. To advance haptic research closer to this goal, this study examines haptic user behavior within the Heat Temperature Module, a haptically enhanced science learning system developed by the HCI research team at the University of Arkansas.

2 Heat Temperature Module

The Heat Temperature Module (HTM) is a science learning haptic application developed at the HCI lab at the University of Arkansas under the supervision of Dr. Chang S. Nam. HTM supports Viscosity Mode, Structure Mode, and Speed Mode. These aptly named modes allows users to haptically explore the Viscosity, Molecular Structure, and Molecular Speed of substances in cold, warm, and hot temperatures. Currently HTM includes the substances CO₂, CS₂, BF₃, SO₃, H₂O, NO₂, NH₃, CH₄, and PCL₃. HTM supports the Novint Falcon Haptic Device as well as a custom Thermal Device developed by the HCI research team at the University of Arkansas (Fig. 1).



Fig. 1. Novint Falcon (left) and HCI Thermal Device (right)

The haptic device and thermal device are used in tandem to allow users to haptically feel an interactive object with one hand, while feeling the corresponding temperature of the object with the other hand. The Novint Falcon is a 2 DOF haptic device capable of rendering haptic forces just over 2 lbs (Novint Falcon, 2010). The HCI Thermal Device is an electronic device that consists of a touchable surface that is capable of rendering controlled temperatures between 0 and 40 degrees Celsius with a standard deviation of 1-2 degrees Celsius.

2.1 Viscosity Mode

Only HTM's Viscosity Mode is considered in this research. This mode contains three different 2D interfaces corresponding to the viscosity of a substance, in a liquid state, at hot (low viscosity), warm (medium viscosity), and cold (high viscosity) temperatures (Fig. 2).

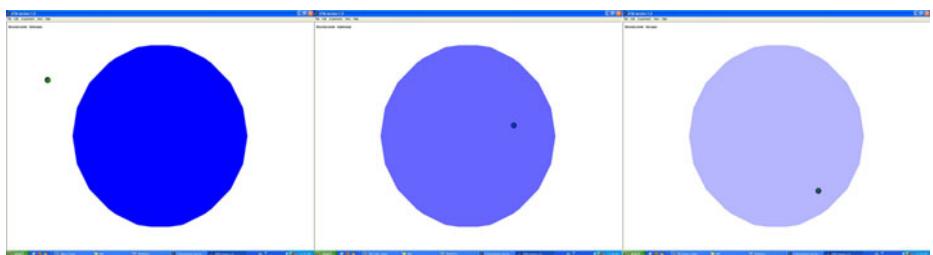


Fig. 2. HTM Viscosity Mode Cold (left), Warm (mid) and Hot (right) interfaces

Viscosity describes how easily a liquid pours or drains through a hole. Each interface contains a single substance, represented by a large two dimensional sphere, centered within a haptic boundary. The sphere haptically exhibits high, medium, and low friction to represent cold, warm, and hot temperature, respectively. Visually, the transparency of the substance changes at each temperature.

2.1 HTM Viscosity Mode Example User Scenario

Within HTM's Viscosity Mode, users simultaneously utilize both hands, one to haptically navigate the virtual environment and explore and interact with haptic objects using the Novint Falcon haptic device, while the other hand feels the corresponding temperature of the haptic environment as well as the haptic objects with the virtual environment using the thermal device. In a typical scenario, users navigate the Viscosity Mode interface, interacting with the substance (i.e. haptic sphere), exploring the area surrounding the haptic sphere, and interacting with the haptic boundary around the virtual environment. To the user, the area surrounding the haptic sphere exhibits no haptic feedback – a haptic sensation similar to effortlessly waving one's hand through the air. Likewise, users are able to move through the inside of the haptic object, feeling the sensation of friction, or a resistant movement, relative to the temperature of the substance – a haptic sensation similar to moving one's hand through thick molasses (Viscosity Mode: cold), water (Viscosity Mode: warm), and steam (Viscosity Mode: hot).

3 Method

3.1 Participants

9 participants were recruited from Kansas State School for the Blind in Kansas City, Kansas. Participants were recruited by instructors at the school. There were six female and seven male participants whose mean (M) age was 13 years (Standard Deviation, SD = 2.30). None of the participants had participated in any previous haptic research conducted by the University of Arkansas.

3.2 Apparatus

This study utilized a Dell PC with a 3.4 GHz Pentium R and 1.0GB of RAM. Two haptic devices were used for Training Exercises and Heat Temperature Module (HTM) tasks: a Novint Falcon haptic device and a Thermal Device designed by the HCI research lab at the University of Arkansas. The Novint Falcon was securely fastened to the desktop using adhesive Duck Tape. The Heat Temperature Module utilizes Adobe Flash CS3 software for visual rendering and a C/C++ Novint SDK for haptic rendering. Haptic features were multithreaded, graphics were rendered at 60 Hz, and haptic control was updated at a rate of 1 kHz.

HTM's Viscosity Mode was utilized to collect and analyze user behavior data. Three unique user interfaces were designed for HTM Viscosity Mode, each with different amounts of applied haptic friction feedback and thermal feedback. Table 1 contains a detailed description of HTM's Viscosity Mode.

Multiple surveys and questionnaires were used to assess haptic and thermal recognition, cognitive workload, and user preference. For haptic and thermal recognition, a 5-point Likert scale questionnaire was utilized (e.g. "How difficult was it to move through the inside of the object from 1 [Easy] to 5 [Difficult]?). For cognitive workload, a NASA TLX was utilized which contains six sub-scales measuring mental demands, physical demands, temporal demands, performance, effort, and frustration. All items were rated on a 10-point Likert scale.

3.3 Procedure

Prior to the experiment, each participant was required to listen to, and agree to an Informed Assent Form to participate in the study. Next, each participant was required to complete a Demographics Form. Then, each participant was required to complete a three-part Training Session. Because many participants were unfamiliar with haptic virtual environments, haptic devices, and Thermal Devices, a Training Session was developed to enable users with a foundational understanding and sensibility of haptic virtual environments, the Novint Falcon haptic device, and the HCI Thermal Device. Participants were allowed to revisit or repeat any part of training they were uncomfortable with or unsuccessful at. Once the tester felt that the participant had achieved sufficient training and had successfully completed all preliminary questions and tasks, the participant was allowed to proceed past training. It should be noted that no visualization was provided for any participant – only the experimenter was allowed to watch the visualization of each experiment on an external monitor.

Once a participant successfully completed the Haptic Training Program, three unique interfaces from the Viscosity Mode key task were presented. The sequence of user interfaces were counterbalanced to remove the influence of the learning process as much as possible. Each interface scenario was conducted as follows: A participant listened to a scenario description outlining the interface and goal(s). During this time, a participant could ask any relevant questions – as long as it did not reveal sensitive information regarding how to go about completing the scenario goal(s). Participants were told to perform each task until the time expired. Upon each scenario's conclusion, the participant was asked to complete a NASA TLX cognitive workload questionnaire. Upon each key task's conclusion, a participant was asked to complete a Key Task Questionnaire to obtain user preference and comments in regards to each interface type within the key task. It should be noted that no visualization was provided for any participant – only the experimenter was allowed to watch the visualization of each scenario on an external monitor. Table 1 provides a detailed analysis of the Viscosity Mode key task.

Table 1. HTM Viscosity Mode key task

Viscosity Mode Task Scenario Time: 30 seconds	Description	There is only 1 subject in the box. Please continue moving left to right, then right to left, back and forth until the time expires. Make sure to touch the wall of the box before you change directions. Your goal is to determine how difficult it is to move through the inside of the object and what temperature the object is.
	Scenario Questions	<ul style="list-style-type: none"> - Please rate how difficult it is to move through the object from 1 (Easy) to 5 (Difficult). - What was the temperature of the object: Cold, Warm, or Hot? - What did the object feel like?
	Purpose	To navigate and locate an object within the haptic boundary and to distinguish different object viscosities and temperatures.

3.3 User Behavior Measurements

A User Behavior Tracking System (UBTS) was developed in order to collect and store all user cursor behavior throughout each scenario of the key task. The UBTS internally stores Novint Falcon position data (Pos X, Pos Y) at an interval of approximately 20 ms. Each position is marked with a timestamp. Upon a scenario's conclusion, HTM outputs a data file containing the results of the UBTS data.

4 Results

Viscosity Mode user behavior data was examined for each participant and each scenario, a UBTS file contained sample size ($n = 1206$) data points. As illustrated in Fig. 3, each Viscosity Mode interface consists of two states: 0 (outside the interactive haptic object) and 1 (inside the interactive haptic object).

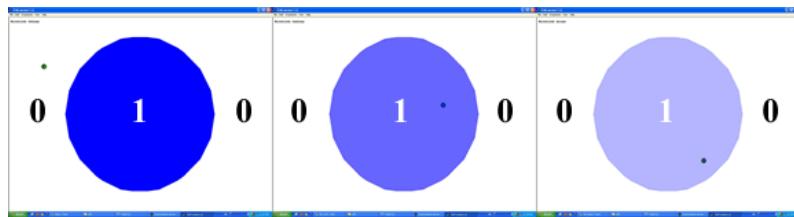


Fig. 3. Viscosity Mode interface state definition

User behavior proportions, per interface, were quantified in Table 2.

Table 2. HTM User Behavior Data Proportions for State 1

Participant	User Behavior State 1 Data Proportions		
	Interface 1 (Cold)	Interface 2 (Warm)	Interface 3 (Hot)
1	60.86%	63.43%	32.42%
2	87.89%	58.79%	57.55%
3	35.57%	28.44%	35.16%
4	26.37%	23.80%	11.19%
5	47.18%	46.43%	27.20%
6	36.07%	18.49%	12.69%
7	53.65%	53.15%	53.07%
8	15.17%	15.09%	0.08%
9	12.94%	2.57%	5.22%

An ANOVA test was conducted to determine significance between Viscosity Mode interfaces for the 9 participants utilizing the user behavior data proportions. Table 3 contains the results of the ANOVA test.

Table 3. User Behavior ANOVA Test Results

F(2,16)	p	Interface 1		Interface 2		Interface 3	
		M	SD	M	SD	M	SD
8.860	0.003	0.418	0.237	0.345	0.216	0.261	0.205

Post-hoc analysis indicated that Interfaces 1 (Cold) and 3 (Hot) were significantly different ($F_{1,8} = 17.7$, $p = 0.0007$) as well as Interfaces 2 (Warm) and 3 (Hot) were significantly different ($F_{1,8} = 5.07$, $p = 0.0387$). However, Interfaces 1 and 2 were not significantly different ($F_{1,8} = 3.82$, $p = 0.0684$).

5 Discussion

This study showed that differences in the amount of haptic feedback result in significant differences in user behavior. In HTM's Viscosity Mode, the amount of haptic feedback, or friction, decreases from Interface 1 (Cold) to Interface 3 (Hot). Likewise, as seen in Table 3, user behavior proportion means (μ) decrease from Interface 1 to Interface 3. The initial explanation can be correlated to Viscosity Mode's haptic feedback (i.e. friction).

As a user moves through the inside of the haptic object, the amount of friction would directly affect how quickly the user's haptic device can move. Therefore, one can reason that higher amounts of friction would result in the user moving more slowly through the haptic object – leading to higher proportions of haptic interaction (State 1). Likewise, lower amounts of friction would result in the user moving more quickly through the haptic object – leading to lower proportions of haptic interaction (State 0). Therefore, it can be reasoned that stronger levels of friction feedback results in higher interactive user behavior proportions. As a real world example, imagine dragging your hand - for the same distance - through honey versus effortlessly moving your hand through air; your hand would remain in the honey for a greater proportion of time due to the high amount of friction applied to the your hand as it moves.

Results also indicated that Interface 1 and 2 were not significantly different, however; Interface 1 and Interface 2 were significantly different from Interface 3. Of the three interfaces, Interface 3 has the lowest level of friction, a sensation akin to saving your hand through water vapor. It is possible that minimal thresholds of friction haptic feedback can be established for a desired level of minimum user interaction data proportions. To establish such desired user behavior thresholds would require further research.

6 Conclusion

This study examined haptic user behavior within the Heat Temperature Module, a haptically enhanced science learning system for learners with severe visual impairments. The overall goal of this research was to advance haptic user interface research in the goal of maximizing ease of use in the domains of haptic interaction and haptic user interface design for both sighted and visually impaired users. Results indicated that higher levels of haptic friction feedback result in higher user interaction proportions of data. Results also suggested that friction feedback thresholds can be determined to establish minimum user interaction data proportions with haptic interface objects, however; more research is needed to establish such thresholds.

References

1. Basdogan, C., Ho, C., Srinivasan, M.A.: Virtual environments for medical training: Graphical and haptic simulation of laparoscopic common bile duct exploration. *IEEE/ASME Transactions on Mechatronics* 3(6), 269–285 (2001)
2. Brewster, S.: The impact of haptic touching' technology on cultural applications. In: EVA Conference Proceedings, pp. 1–14 (2000)

3. Brooks, F.P., Ouh-Young, M., Batter, J., Kilpatrick, P.: Project GROPE - Haptic displays for Scientific Visualization. Computer Graphics (Proc. SIGGRAPH 1990), pp. 177-185 (1990)
4. Burdea, G.C.: Haptic Sensing and Control. In: Force And Touch Feedback For Virtual Reality, pp. 13–40. John Wiley & Sons, New York (1996)
5. Christensson, J.: Individualising Surgical Training in a Virtual Haptic Environment - Design of flexible teaching strategies to support individuals (Master's thesis). IT University of Göteborg, Göteborg, Sweden (2005)
6. Christodoulou, S.P., Garyfallidou, D.M., Gavala, M.N., Ioannidis, G.S., Papatheodorou, T.S., Stathi, E.A.: Haptic devices in Virtual Reality used for Education: Designing and educational testing of an innovative system. In: Proceedings of International Conference ICL 2005 (Interactive Computer Aided Learning), Villach, Austria (2005)
7. Gunther, E., O'Modhrain, S.: Cutaneous Grooves: Composing for the Sense of Touch. *Journal of New Music Research* 32(4), 369–381 (2003)
8. Hubbold, R.: Collaborative stretcher carrying: A case study. In: Proceedings of 2002 EUROGRAPHICS Workshop on Virtual Environments, pp. 30–31 (2002)
9. Ino, S., Shimizu, S., Odagawa, T., Sato, M., Takahashi, M., Izumi, T., Ifukube, T.: A Tactile Display for Presenting Quality of Materials by Changing the Temperature of Skin Surfaces. In: Proc. IEEE 2nd International Workshop on Robot and Human Communication, pp. 220–224 (1993)
10. Jones, M.G., Minogue, J., Treter, T.R., Atsuko, N., Taylor, R.: Haptic augmentation of science instruction: does touch matter? *Science Education* 90, 111–123 (2006)
11. Kilpatrick, P.: The use of kinesthetic supplement in an interactive system (Ph.D. thesis). University of North Carolina, Chapel Hill, NC (1976)
12. MacLean, K.E.: Designing with haptic feedback. In: Proceedings of the IEEE International Conference on Robotics and Automation, pp. 783–788 (2002)
13. McLinden, M.: Haptic exploratory strategies and children who are blind and have additional disabilities. *Journal of Visual Impairment and Blindness*, 99–115 (2004)
14. Miller, T., Zeleznik, R.C.: The design of 3D haptic widgets. In: Proceedings 1999 Symposium on Interactive 3D Graphics, pp. 97–102 (1999)
15. Millman, P.: Haptic Perception of Localized Features (Ph.D. thesis). Northwestern University, Evanston, IL (1995)
16. Minogue, J., Gail Jones, M., Broadwell, B., Oppewall, T.: The impact of haptic augmentation on middle school students' conceptions of the animal cell. *Virtual Reality* 10(3), 293–305 (2006)
17. Falcon, N.: Technical Specs (2010),
http://home.novint.com/products/technical_specs.php
18. Paulu, N., Martin, M.: Helping your child learn science (Technical Report ED331727). US Department of Education (1991)
19. Roussou, M.: Learning by doing and learning through play: an exploration of activity in virtual environments for children. *ACM Computers in Entertainment* 2(1) (2004)
20. Ruffaldi, E., Morris, D., et al.: Standardized evaluation of haptic rendering systems. *Haptic Interfaces for Virtual Environment and Teleoperator Systems*. In: IEEE VR, pp. 225–232 (2006)
21. Sallnas, E.L., Rassmus-Grohn, K., Sjostrom, C.: Supporting presence in collaborative environments by haptic force feedback. *ACM Transaction on Computer-Human Interaction*, 461–476 (2000)

22. Samur, E., Want, F., Spaetler, U., Bleuler, H.: Generic and systematic Evaluation of Haptic Interfaces Based on Testbeds. In: Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems, San Diego, CA (2007)
23. Sauer, C.M., Hastings, W.A., Okamura, A.M.: Virtual environment for exploring atomic bonding. In: Proceeding of Eurohaptics, pp. 232–239 (2004)
24. Swindells, C., Maksakov, E., MacLean, K.E., Chung, V.: The role of prototyping tools for haptic behavior design. In: IEEE Symposium on Haptic Interfaces for Virtual Environments and Teleoperator Systems (HAPTICS 2006), Washington, DC (2006)
25. Yu, W., Brewster, S.: Evaluation of multimodal graphs for blind people. Universal Access in the Information Society, 105–124 (2003)

Usability Testing of the Interaction of Novices with a Multi-touch Table in Semi Public Space

Markus Jokisch, Thomas Bartoschek, and Angela Schwering

Institute for Geoinformatics, University of Münster, Weselerstraße 253,
48151 Münster, Germany

{markus.jokisch,bartoschek,schwering}@uni-muenster.de

Abstract. Touch-sensitive devices are becoming more and more common. Many people use touch interaction, especially on handheld devices like iPhones or other mobile phones. But the question is, do people really understand the different gestures, i.e., do they know which gesture is the correct one for the intended action and do they know how to transfer the gestures to bigger devices and surfaces? This paper reports the results of usability tests which were carried out in semi public space to explore peoples' ability to find gestures to navigate on a virtual globe. The globe is presented on a multi-touch-table. Furthermore, the study investigated which additional gestures people use intuitively as compared to the ones which are implemented.

1 Introduction

Multi-touch is used more and more on many different devices and in many places, especially on handhelds but also on walls or tables in (semi) public space. Many people of all ages and experiences have access to multi-touch-devices. But it is not clear if all of them are familiar with ways to interact with such devices. Although many people commonly use virtual globes on their PCs to navigate the whole world, it is difficult for them to use a virtual globe on a multi-touch-table. This usability tests examined how a wide audience of untrained users in semi public space deal with a multi-touch-table on which a virtual globe is shown. We conducted the usability tests to introduce another form of human computer interaction. Especially for people who are not familiar with technical issues the multi-touch-devices could help them to interact with a computer. One could conjecture, the interaction is easier for them because it is very direct and they need not use other input devices such as mouse and keyboard. Hence, it is very important that the system works as they suggest and the gestures must be as intuitive as possible. Obviously, we chose novice users to do the usability tests because they are not familiar with other multi-touch-devices. We postulated that they do not know the gestures which are implemented in other systems nor could they infer gestures on our multi-touch-table from their previous knowledge. Because of this, they are most comparable to the mentioned user group which is not familiar with technical issues at all. We carried out our test not as a laboratory experiment but as a field study, because in this case we could ensure that the surroundings and the environment are similar as possible to the situation where such tables would be placed for the final usage.

The used Software is a GoogleEarth multi-touch-version¹. The main goal of the research is to examine if users of the table can find and use the gestures which are necessary for the use of the multi-touch table without any explanation. The following gestures are examined: One gesture to move the map (pan), one to enlarge or shrink the image (zoom) and one to change the inclination angle of the “camera” (tilt). These are the three basic operations which can be performed on a virtual globe. Furthermore, we investigated whether users perform other gestures to fulfill the task and how these gestures look like.

The first main test took place at the institute of political science of the University of Münster. In both tests the subjects were asked to fulfill a task on the multi-touch-table in which every gesture has to be executed. To determine the goal the user had while making a gesture, users were asked to talk about it.

In a second usability test, we investigated another group of possibly novice users: Children of a primary school in Rheine (Westphalia, Germany). We repeated our study with children, because we expected the children to be real novices with no previous experience in multi-touch. Furthermore, in pilot studies multi-touch-tables have been successfully used as learning devices in schools. The setting of the usability test had to be slightly adapted from the one with the adult user to meet the children’s and the school’s requirements.

Over the course of the paper the following issues are explained in detail: The second section introduces related usability studies with touch-sensitive devices in different settings. In the third section, we describe the technical background of our own study and the implemented gestures are explained. Furthermore, we report on our test results in the fourth section. Finally, we conclude with a discussion and give directions for future work.

2 Related Work

The usability of multi-touch-devices has been investigated in several studies previously. While our study was located in semi public space, many of the other usability experiments have been conducted as laboratory tests, such as Moscovich’s et al. [9] and Wu’s et al. [18]. In Moscovich’s test, subjects could drag and manipulate easy geometric objects like squares and circles with one or two fingers. In Wu’s test the subjects had to plan a room on a multi-touch-table with a special room planning software. To fulfill the task the subjects had to use gestures, but they were explained to the subjects before the test started. Only a few of the tests refer to virtual globes like they have been used in our research, e.g. the one of Tse et al. [16], which was rather an informal test with colleagues during the implementation of a set of gesture for navigating on virtual globes. Furthermore, the gesture set was complemented by speech recognition. In this test, panning and zooming, which have been the same as in this test, could be found easily. The expected tilt gesture, an upwards movement of the hand, a 3D gesture, could not be implemented on the used hardware. The implemented gesture was the usage of five fingers but this one was hard to find for the subjects.

¹ <http://nuigroup.com/forums/viewthread/5422/>

Other usability tests took place in public places as in this study, however in very different environments. One of them is conducted in the city center of Helsinki, Finland. A multi-touch-wall was installed by Peltonen et al. on which you can manipulate pictures from the Internet platform “Flickr” by using gestures. They mostly studied the group behavior of the users, so called multi-user approach and not a multi-touch single-user approach as we do. Furthermore, they studied the usage of gestures but this was done during the manipulation of pictures and not of virtual globes [10]. Also in a city center, Schöning et al. asked users of a multi-touch-wall about their behavior while using virtual globes. In this study emphasis was not put on the usage of the gestures although they were needed to fulfill the given task [15].

In exhibitions, field studies with multi-touch-tables and walls have been carried out already. During the exhibition “Deutschland, Land der Ideen” (“Germany, Land of ideas”) pedestrians were observed in a pedestrian underway in the city of Münster (Westphalia), Germany, while using a multi-touch-wall. The focus was on the spontaneous interaction with the shown virtual globe and the group behavior of the users and not on the used gestures [3]. In the “Museum of National History” of Berlin, Hornecker installed a multi-touch-table with so called “Tree of Life”-Software and observed how users approach the table and use the software which did not provide a virtual globe or a map. The shown content mostly was in question/answer form [7].

In all these works the interaction with the multi-touch-device is possible through various gestures but testing the ability especially of novice users to find and use them is very rare, e.g., Daiber et al. and Hornecker mention it in passing [3, 7].

Possible intuitive gestures were only investigated by using questionnaires or laboratory tests, but not by the use of the multi-touch-device on site [4, 17]. Only Daiber et al. ask for gestures to use virtual globes [3]. Hornecker describes gestures which are made in the situation of the usage but they are not used to navigate on a virtual globe [7].

With Children a few studies have been done, but none with virtual globes. Most of these studies analyzed the group behavior of the subjects and the cooperative usage of such a device [2, 11]. A usability test on a multi-touch-device was only done by Rick et al. [13] and Harris et al. [6]. They used learning software for children but mostly studied the differences between a single- and a multi-touch version of their software. Furthermore no gestures had to be used and the usage was explained to the children before. Many studies have searched for the best method to study children. They compared questionnaires, interviews and different “Thinking-Aloud”-Tests [5]. The result was that “Thinking-Aloud”-Tests and especially the “Constructive Interaction” Method, where subjects do a “Thinking-Aloud”-Tests in small groups, are the best methods [1]. Furthermore, the studies showed that questionnaires must not be too difficult and not have too many answering options [5]. Another point is that fun is important point for children in their rating [8]. Their findings strongly influenced our design of the usability test with children.

3 Fundamentals

3.1 Technical Background

The software used was a Google Earth plug-in for web browsers, which allows multi-touch-interaction with Google Earth. The software which detects the contact with the

surface is Community Core Vision², an open source software of the NUIGroup. The hardware is a self-made multi-touch-table in which the technical equipment, like the used PC, the projector and infrared lights, is placed. The table has a width of 110cm, a depth of 60cm and a height of 79cm. The size of the display is 84x52 cm. To recognize the touches on the surface the rear diffused illumination technique is used. In this method the embedded camera recognizes the infrared light which is reflected by the touches on the surface in addition to the standard reflection of the frame.

3.2 The Gestures

The gesture to pan the map is done with one finger. The finger is placed on the surface, the map is anchored on this point and is panned when the finger moves.

There are two ways of performing the zoom gesture: With one hand and two of its fingers or with two hands and one finger of each. In both cases there are two points of contact with the surface. The two fingers are moved together to zoom in whereas they are moved apart to zoom out. In general this gesture is called “pinch to zoom”.

The gesture to tilt the camera perspective, e.g., to show differences in the terrain or 3D-buildings is a gesture with three points of contact. It can also be performed with one or two hands. Two contact points are not moved and the third is moved in vertical direction. With a contact from the front side of the table to the rear side the inclination of the camera becomes more horizontal and the contrary if the finger is moved from the rear side to the front side.

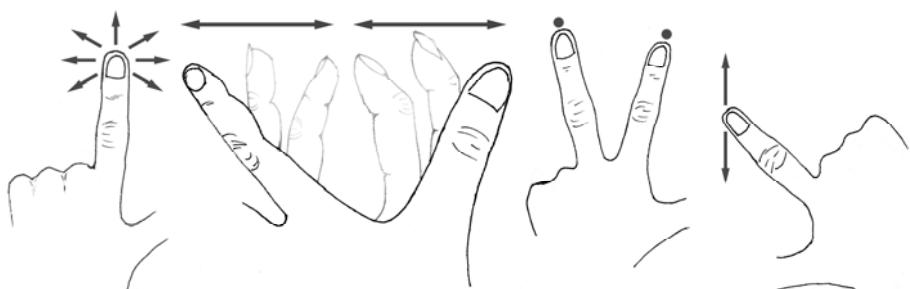


Fig. 1. The investigated gestures pan, zoom and tilt

4 The Tests

4.1 Main Test with Adult Subjects

The main-test took place in the lobby of the building of the institute for political science at University of Münster. The subjects were students with no background in computer science. The subjects were asked whether they want to fulfill a short task at the multi-touch-table right after their arrival. They were given the following task: “Please navigate to a point of interest, show it in detail and tilt the camera to see if there are 3D-buildings.”

² <http://ccv.nuigroup.com/>

It was not only observed whether the subjects found the gestures to operate the table and which ones they found but also how many tries the subjects needed to find it. We asked the subjects whether they found it easy to apply the gesture and whether they had problems to find the gesture.

While the subjects were doing the task they were observed by the interviewer whether they could find and use the implemented gestures. After the test the interviewer asked some statistical questions about age and gender followed by questions about the subject's experiences and feelings in the test. In addition video recordings were made to verify the found gestures and to recognize the gestures made intuitively in an analysis after the test.

In the case where a subject did not find a gesture the interviewer did not interrupt the subject. Most subjects stopped trying after three or four unsuccessful attempts. However, the right gesture was explained to the subject in two cases. On the one hand, if the subject asked for the right gesture after some time or on the other hand if the subject was thinking for about 30 seconds without operating the table.

A total of sixteen subjects participated in the test. Ten of them were female, six male. A female subject had to be excluded from the analysis, since she said to have quite a bit experience with multi-touch and could not be seen as a novice user. Her results have not been analyzed. Thirteen subjects were right-handed, two were left-handed and all of them were between 20 and 30 years old. Everyone knew the Google Earth software. Only two subjects said to have medium experience with multi-touch. Only 22.2% of experienced users and 13.3% in general had experiences with a large multi-touch-device.

73.3% of the users approached at the table when it was not in use while the other 26.7% entered it when somebody was using it. So it is possible that these subjects learned from the user who was at the table previously [14]. Every user found the gestures to pan. The zooming gesture was found 80%. The tilt gesture was only found 6.7%.

Seven subjects made unsuccessful attempts until they found the panning gesture. But everyone only had one unsuccessful attempt until finding the gesture. A remarkable fact is that some of them changed back to the wrong gesture although they had found the right one. The average of unsuccessful attempts was 0.47. Multiple uses of the same wrong gesture were only counted as one unsuccessful attempt.

While searching the zoom gesture, the subjects made more unsuccessful attempts. Seven subjects made one, one subject made two and two subjects made three unsuccessful attempts. On average the subjects made 1.25 unsuccessful attempts with a standard deviation of 0.92. The subjects that did not find the right gesture of course made unsuccessful attempts only. Also in this case every usage of the same wrong gestures was only counted once.

The subject who found the tilt gesture made one wrong gesture until he found the right one. Furthermore only two attempts to find the gesture were made. All other subjects said to have no idea how it could work and did not try to find out.

Also the duration until the subjects found the gesture was measured. It differs in the case of the panning gesture from one up to 33 seconds and for the zoom gesture from five up to 75 seconds.

Subjects that had medium experience made no mistake during the usage of the pan gesture. The number of unsuccessful attempts aggregates to seven, divided nearly similar to the other two groups.

Three subjects that almost had no experience with multi-touch were able to find the zoom gesture. But they had, on average, 2.33 unsuccessful attempts until they found it. This value is more than one attempt below general average. Subjects with a little experience or medium experience only had 0.86 and 0.5 unsuccessful attempts respectively. So absolute novice users make more mistakes but half of them were able to find the zoom gesture.

Also the assessment of the simplicity was different in these user groups. Medium experienced subjects described the simplicity of finding the gesture easy or very easy, little experienced users said it was neither easy nor difficult. In both cases this is above general average. Only subjects with almost no experience assessed it on average as difficult. The results in simplicity of the usage of the gestures are similar. Again the medium experienced subjects assessed the gestures on average as easy to use, the subjects with little experience again said it was neither difficult nor easy, and although those with almost no experience gave the worst marks, they were still passable. But in this case one has to keep in mind that they only evaluate the gestures they have used and half of the subjects only found and used the zooming gesture.

The average duration until the subjects found the panning gesture was 9.6 seconds and the zoom gesture 29.1 seconds. In both cases the medium experienced were the fastest with 1.5 and 8.5 seconds respectively. The second fastest in finding the panning gesture were the inexperienced with 7.5 seconds and third were the little experienced with 13.7 seconds. The reason for this surprising result could be that the inexperienced touched the table much more carefully and so did not use the whole for panning first. The results for the zoom gesture were 8.5, 24.6 and 53.3 seconds in the expected order.

The Spearman Rho correlation analysis shows that the experience of users is strongly correlated to the success of finding and using the gestures. Also these two questions are strongly correlated to each other. Furthermore strongly correlated are the number of mistakes and the duration until the finding of the zoom gesture. This variable is also strongly correlated to the two simplicity questions. Not as strongly connected as the just mentioned variables, are the experience to the mistakes and duration of the zoom gesture. Other variables are not correlated.

4.2 Main Test with Children

Following previous studies [1, 5, 8] mentioned already in the related work, we adapted our study design for the main test with children. We invited children in small groups to test the multi-touch-table. Besides the constructive interaction method, a short questionnaire was used. The children were asked about their experience with computers and especially multi-touch. Furthermore, they were asked if they had fun in using the table and whether they found it easy or difficult. All these questions were only separated in two answer possibilities but in the simplicity question the children gave three different answers. Additionally, statistical notices about the grade, the gender and the found gestures of the children were made. Because of the group situation it was too difficult to count mistakes while searching the gestures. But the duration until a group found the right gesture was measured.

42 children in nine groups took part in the test. 23 were female and 19 male, 36 right-handed and six left-handed. All of them had experiences with computers and had fun in fulfilling the exercise.

20 children said they had no experience with multi-touch, 21 said they had experience, one said "a little". The devices on which the children had collected their experiences were very different and not all are comparable to the used multi-touch-table. The simplicity was judged easy by seven subjects, "middle" by 26 subjects and difficult by eight subjects. One child made no statement.

In all groups the panning gesture was found by at least one child, all together by 25. The zooming gesture was found by three experienced children in three different groups, who had different experience on average. Nobody found the panning gesture and not many attempts were made to find it. The duration until the gestures were found in the groups differs between 10 and 185 seconds for the panning gesture and between 48 and 195 seconds for the zooming gesture.

Experienced children judged the table a bit better than the inexperienced but if you do a correlation analysis between these two variables you can see that they are not significantly connected. This counts for the cases subjects and groups. Also the duration until the finding of the gestures and if the zoom gesture was found is not connected to any other variable.

So it can be said that for children it seems not to be very important as for adult subjects to have experiences to work with the table properly, although all children who found the zoom gesture had experiences.

4.3 Results and Analysis of Additionally Used Gestures

For the analysis of additionally used gestures we included all subjects in the analysis, i.e., the subjects of our pretest, main test and those who were excluded from the analysis due to their previous knowledge or since they did not agree to participate but still used the table.

Our analysis of additional gestures showed that many users prefer to pan the map by using multiple fingers instead of one finger (seventeen (adult), twenty (children)). Moreover, somebody tried to rotate the globe by snapping his finger over the surface and it was tried to move the globe itself when the zoom level was low and not the whole display was covered.

When zooming, there were two gestures that have been mainly used in addition to the implemented ones. This is the known double click of the desktop version (eighteen, fifteen), and the use of all fingers of the hand in opposite to the thumb instead of just one finger (thirteen, twelve). Two users who made the gesture implemented it as a two-hand gesture, but not with one finger per hand, but with two fingers. In addition one subject attempted to click on the "stars" which mark the towns in GoogleEarth, another one attempted to press on the surface for a longer time and as well people searched for buttons and commands in the menu bar on the right hand side of the display.

For the tilt of the camera mainly the whole hand or several fingers were tipped over the surface (three). Moreover many gestures were used only once or twice: The turn of all fingers on the surface, a sliding movement with more than one finger and the back of the hand. Furthermore, the implemented gesture was one time used "inverse", i.e., one finger attached and two fingers moved. Even for this gesture buttons were searched.

In particular, for panning and zooming there were gestures which have been used intuitively very often. An implementation of these would simplify the handling of the table. Especially the gesture double-clicking or double-touching in this case and pinch to zoom with more than one finger for zooming, and the possibility to move the map with more than one finger for panning would be helpful. Based on our usability test we could not find a set of intuitive gestures for tilting and suggest providing users with additional information on this functionality. We believe that the reasons for the usability problems are caused by the dimensionality: Tilting of the map is a 3D action, but the surface allows only for 2D gestures.

For the tilt gesture it is notable that many users of the table, who have not been given the task, were not even aware that there is the possibility of tilting the image. They only noticed it when they had accidentally tilted the image. But after that they did not try to restore it but left the table quickly or pressed the “Reset Position”-button.

In the children test we found similar results with respect to additionally used gestures. However, interesting differences were discovered regarding the behavior of children: Children tried out many more different gestures per subject, except for the tilting operation. For this operation the most used gesture was the sliding of a flat hand over the surface. While adults spent more time on thinking, children were quicker to test numerous different gestures. Children obviously had fewer reservations to touch the table (however this might be influenced by the slightly different group setting). Another interesting finding is that children tried out the same gestures not only for one but for different operations. These were sliding of the whole hand over the table, the rotation of the fingers on the table and a parallel sliding of some fingers from different hand over the table. All three were tried for zooming and tilting. A correlation analysis of the experience of the groups and their additionally used gestures showed that more experienced children tried more gestures than inexperienced ones. The reason of this result could be seen very well throughout the tests. Inexperienced children were more afraid to touch the table than experienced ones and so they did fewer gestures. The children’s judgment of the simplicity is not correlated to the number of additionally used gestures.

5 Conclusion and Future Work

The tests show that many but not all young people aged between 20 and 30 years know intuitively how to operate a multi-touch-table even if they have no experiences. Especially panning and zooming is no problem for most of them. They find the needed gestures and are able to use them. Some other gestures like the double click or the panning with more than one finger which have been done by some subjects should be implemented and tested in future.

Only some of the subjects were able to find the tilt gesture. So this one does not seem to be intuitive but our test results did not reveal other alternative gestures which could be implemented instead. In further developments the gesture should not be used without an explanation.

Although children interact faster and more frequently with the table, it is more difficult for them to find the right gestures and they should be provided with more explanations. It does not matter if the children have a bit of experience or not.

The present study tested users who are up to 30 years old. In future work older user groups, so called digital immigrants [12], should be studied. In addition, people without experiences with multi-touch and without experiences with GoogleEarth should be studied. In another study of children the groups should be smaller or at least only one child should be studied because too many hands are confusing for the table and the study as a whole.

Acknowledgements. Thanks to Henni for co-reading the paper, her help in design issues and her mental support, to Swantje, Lutz, Imad and Malumbo for their support in logistical questions and to Andres and Wadim for their Soft- and Hardware support. And of course we want to thank all participants of the tests.

References

1. Als, B.S., Jensen, J.J., Skov, M.B.: Comparison of Think-Aloud and Constructive Interaction in Usability Testing with Children. In: Proceedings of the 2005 Conference on Interaction Design and Children, Boulder, pp. 9–16 (2005)
2. Browne, H., Bederson, B., Druin, A., Sherman, L.: Designing a Collaborative Finger Painting Application for Children. College Park (2000)
3. Daiber, F., Schöning, J., Krüger, A.: Whole Body Interaction with Geospatial Data. In: Butz, A., Fisher, B., Christie, M., Krüger, A., Olivier, P., Therón, R. (eds.) SG 2009. LNCS, vol. 5531, pp. 81–92. Springer, Heidelberg (2009)
4. Epps, J., Lichman, S., Wu, M.: A study of hand shape use in tabletop gesture interaction. In: Proceedings of the Conference on Human Factors in Computing Systems, pp. 748–753. ACM, Montréal (2006)
5. Donker, A., Markopoulos, P.: A comparison of think-aloud, questionnaires and interviews for testing usability with children. In: Faulkner, X., Finlay, J., Détienne, F. (Hrsg.) People and Computers XVI: Memorable Yet Invisible, pp. 305–316 (2002)
6. 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, Rhodes 2009, vol. 1, pp. 335–344 (2009)
7. Hornecker, E.: I don’t understand it either, but it is cool – Visitor Interactions with a Multi-Touch Table in a Museum. In: Proceedings of the 3rd IEEE International Workshop on Horizontal Interactive Human Computer, pp. 121–128. IEEE, Amsterdam (2008)
8. Markopoulos, P., Bekker, M.: On the assessment of usability testing methods for children. *Interacting with Computers* 15(2), 227–243 (2003)
9. Moscovich, T., Hughes, J.F.: Indirect mappings of multi-touch input using one and two hands. In: Proceedings of the Twenty-sixth Annual SIGCHI Conference on Human Factors in Computing Systems, pp. 1275–1284. ACM, Florence (2008)
10. Peltonen, P., Kurvinen, E., Salovaara, A., Jacucci, G., Ilmonen, T., Evans, J., Oulasvirta, A., Saarikko, P.: ‘It’s Mine, Don’t Touch!’: Interactions at a Large Multi-Touch Display in a City Centre. In: Proceedings of the Twenty-sixth Annual SIGCHI Conference on Human Factors in Computing Systems, pp. 1285–1294. ACM, Florence (2008)
11. Plichta, M., Nischt, M., Joost, G., Rohs, M.: Touching Newton: a round multi-touch table for collaborative learning among children, Berlin (2005)
12. Prensky, M.: Digital Natives, Digital Immigrants. *On the Horizon* 9(5), 1–15 (2001)

13. Rick, J., Harris, A., Marshall, P., Fleck, R., Yuill, N., Rogers, Y.: Children designing together on a multi-touch tabletop: an analysis of spatial orientation and user interactions. In: Proceedings of the 8th International Conference on Interaction Design and Children, Como 2009, pp. 106–114 (2009)
14. Russell, D.M., Drews, C., Sue, A.: Social Aspects of Using Large Public Interactive Displays for Collaboration. In: Proceedings of the 4th International Conference on UbiComp 2002, Göteborg, pp. 663–670 (2002)
15. Schöning, J., Hecht, B., Raubal, M., Krüger, A., Marsh, M., Rohs, M.: Improving interaction with virtual globes through spatial thinking: helping users ask “why?”. In: Proceedings of the 13th International Conference on Intelligent User Interfaces, pp. 129–138. ACM, Gran Canaria (2008)
16. Tse, E., Shen, C., Greenberg, S., Forlines, C.: Enabling interaction with single user applications through speech and gestures on a multi-user tabletop. In: Proceedings of the Working Conference on Advanced Visual Interfaces, pp. 336–343. ACM, Venice (2006)
17. Wobbrock, J.O., Morris, M.R., Wilson, A.D.: User-Defined Gestures for Surface Computing. In: Proceedings of the 27th International Conference on Human Factors in Computing Systems, pp. 1083–1092. ACM, Boston (2009)
18. Wu, M., Balakrishnan, R.: Multi-finger and whole hand gestural interaction techniques for multi-user tabletop displays. In: Proceedings of the 16th Annual ACM Symposium on User Interface Software and Technology, pp. 193–202. ACM, Vancouver (2003)

Niboshi for Slate Devices: A Japanese Input Method Using Multi-touch for Slate Devices

Gimpei Kimioka, Buntarou Shizuki, and Jiro Tanaka

Department of Computer Science, University of Tsukuba,
SB1024 University of Tsukuba, 1-1-1 Tennodai, Tsukuba-shi, Ibaraki, Japan
`{gimpei, shizuki, jiro}@iplab.cs.tsukuba.ac.jp`

Abstract. We present Niboshi for slate devices, an input system that utilizes a multi-touch interface. Users hold the device with both hands and use both thumbs to input a character in this system. Niboshi for slate devices has four features that improve the performance of inputting text to slate devices: it has a multi-touch input, enables the device to be firmly held with both hands while text is input, can be used without visual confirmation of the input buttons, and has a large text display area with a small interface. The Niboshi system will enable users to type faster and requires less user attention to typing than existing methods.

1 Introduction

Slate devices, such as the iPad or the Galaxy Tab, are mobile devices that have an approximately 7 by 10 inch touchscreen as the main interface. They have more possible applications and bigger screens than smartphones or PDAs, so they are becoming increasingly widely used. Users tend to input text more on these devices than they do on smartphones or PDAs.

Onscreen QWERTY keyboard (QWERTY) is an existing method commonly used to input Japanese on slate devices. Users input Japanese text by using the alphabet to spell out Hiragana characters phonetically, just like with a hardware-based QWERTY keyboard. For example, 日本 is a kanji word that means Japan, and pronounced as Nippon. This word can be input the pronunciation with Alphabet and converted into kanji afterwards. We believe that this technique is inefficient since it simply mimics a hardware-based QWERTY keyboard and does not consider the characteristics of slate devices. There are three problems with using QWERTY on slate devices.

- Users cannot input text while holding the device firmly. QWERTY is efficient only if the device is put onto a desk or table since all the fingers of both hands are required in order to use it as if it were a hardware-based QWERTY keyboard. There are many situations in which users may not have access to a table when using the device, such as when standing in a train or sitting on a bench. In those cases, users have to hold the device with both hands, or hold it with one hand and use only the other hand to input text. Slate devices are heavier than smartphones or PDAs, which means their stability while they are held is compromised, or only one hand can be used for inputting.

- Users need to look at both the text display area and the input buttons while they are inputting. Since QWERTY does not offer tactile feedback to users when buttons are touched, users have to look at the button they want to push.
- The user interface dominates part of the screen because it has to be big enough to be consistent with the hardware-based QWERTY keyboard. This decreases the visual accessibility of the document when the document is long and cannot be displayed without scrolling.

Since QWERTY is not quite optimized for slate devices as these show, we solved those problems by developing a new Japanese input system specialized for slate devices.

We have been developing a Japanese input system that utilizes a multi-touch interface for smartphones [1]. This system enables users to input a series of Japanese characters with both thumbs simultaneously. In this paper, we present Niboshi for slate devices, which is a Japanese input system modified from our previous work on smartphones.

Niboshi for slate devices uses a multi-touch interface, as our previous system does. The multi-touch interface enables users to input a character as quickly as by one stroke of one finger, once they have been trained in its use. The system also enables characters to be input without the user having to visually confirm the button being pressed because its button layout can be calibrated and the system also provides visual feedback.



Fig. 1. How to hold device in portrait mode (left) and landscape mode (right)



Fig. 2. Interface of Niboshi

2 Related Works

There are related works about text input systems on touch screen devices, kanji conversion system and multi-touch input system. Modified button layouts on touch screen devices are presented in OPTI [2] and Metropolis [3]. Shiohara et al. presented a button layout for Japanese input system that considers kinematic characteristics of thumb movements [4]. This system is about single handed input and our system uses the button layout that can be calibrated for both hands. TagType [5] suggests that

users use both thumbs to input Japanese text. TagType can only handle one input at a time, and our system uses multi-touch. Shin et al. presented a multi-touch Korean text input system that reduces the stroke number of input [6]. This system is for Korean and uses the multi-touch to add the variety of consonants.

3 Niboshi for Slate Devices

Our Japanese input system, Niboshi for slate devices, is described in more detail in this section.

3.1 How to Hold Device

In Niboshi for slate devices (Niboshi), the device is held with both hands. Users can hold the device in portrait or landscape mode (Fig. 1). In portrait mode, more text is visually accessible because the screen can show more lines. In landscape mode, the device can be held more stably because the device's center of gravity is between the hands. Users can simply hold the device as they want, and the system will recognize the orientation and provide the appropriate interface.

3.2 User Interface

The interface of Niboshi is shown in Fig. 2. The buttons are divided into two sets, and each set to appear directly under both thumbs. Consonant buttons are on the left, and vowel buttons are on the right. Function buttons, such as delete button, are placed in relation to the other buttons.

Since users have different sized hands and may hold different parts of the device, the location of the buttons can be calibrated specifically for the user. Niboshi has the calibration function that is explained later in this paper.

3.3 How to Input

The written Japanese language basically consists of three types of character sets: hiragana, katakana, and kanji. Hiragana and katakana are phonetic characters and are phonetically interchangeable. These characters generally start with a consonant and end with a vowel sound. In Niboshi, users can input hiragana and katakana in the same way. They may choose hiragana or katakana before they start inputting. Kanji can be converted from hiragana/katakana depending on how they are pronounced.

Both hiragana and katakana have several types of sounds: voiceless sounds, voiced consonant sounds, p-consonant sounds, contracted sounds, and a double consonant. In Niboshi, voiceless sounds are input in a basic way, and the other sounds can be converted from the voiceless sounds as the diverted sound from voiceless sound.

Voiceless Sounds. For a voiceless sound, a user inputs a hiragana/katakana by choosing a consonant with his/her left thumb and a vowel with his/her right thumb. Character input is confirmed only by the right thumb, so users do not have to cancel an input even if the wrong consonant was chosen. Both hiragana and katakana can be input using the procedure shown below and in Fig. 3.

1. Choose a consonant with your left thumb and press on it. The right-side buttons vary depending on which consonant is being pressed (Fig. 3a).
2. Choose a vowel with your right thumb. The text display field shows the hiragana or katakana corresponding to the combination of the consonant and the vowel you are pressing, so you do not have to visually confirm which buttons you are holding (Fig. 3b).
3. Release your right thumb to confirm the character. The left thumb does not have to be released, which reduces the actions required to continue inputting text (Fig. 3c).
4. Move your left thumb to the consonant button that you want to input next. Repeat the procedure from step 1 (Fig. 3d).

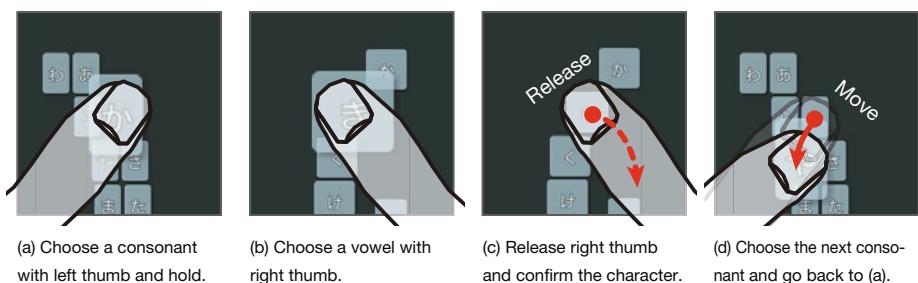


Fig. 3. How to input hiragana/katakana

Table 1 shows what character will be input from which combination of consonant and vowel. All voiceless consonant sounds in Japanese are determined from these combinations.

Table 1. Voiceless sound combinations

		vowel				
		a	i	u	e	o
consonant	a	あ	い	う	え	お
	k	か	き	く	け	こ
	s	さ	し	す	せ	そ
	t	た	ち	つ	て	と
	n	な	に	ぬ	ね	の
	h	は	ひ	ふ	へ	ほ
	m	ま	み	む	め	も
	y	や	(not in use)	ゆ	(not in use)	よ
	r	ら	り	る	れ	ろ
	w	わ	ー	ん	(space)	を

Other Sounds. Sounds other than voiceless sounds are phonetically diverted from voiceless sounds. A gesture is introduced to convert a voiceless sound into a diverted sound. This reduces the number of strokes needed to input a character. The gesture

continuously follows the right thumb touch so that the user can confirm a character and choose the diverted sounds at the same time.

The gesture to convert a sound is performed by the right thumb. The user slides the thumb in the direction of the thumb joint (Fig. 4). Users may also use the MicroRolls [7] gesture (Fig. 4). MicroRolls is a way for users to roll their thumbs up/down on the screen.

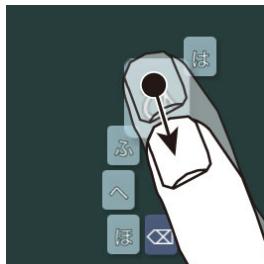


Fig. 4. Gesture for converting sound



Fig. 5. MicroRolls gesture

Kanji Conversion. One of the Japanese language characteristics is that it has a lot of homonyms, and one kanji word cannot be determined phonetically. Many Japanese input systems adopt the way that the system shows several candidates of the kanji word and let the user choose to determine what kanji to input. We also use this technique in Niboshi.

In POBox [8], the system dynamically presents kanji conversion candidates every time a user input a hiragana. In our system, the conversion candidates are shown every time a user input a character like POBox. The system predicts the converted kanji word from what the user has input and show them as the candidates.

Five candidates are shown at a time close to the text input area. The candidates are always shown when the user input, and user can confirm the input whenever he/she wants. To confirm the conversion, the user can hold the conversion button next to consonant buttons, and press the vowel button that corresponds to the candidate he/she wants to choose.

This conversion could be done several times for one input. For example, the ‘ha’ sound has two diversions: ‘ba’ and ‘pa’. A user simply does the gesture back and forth before releasing his/her thumb, and the system switches the diversion every time it recognizes the gesture.



Fig. 6. Kanji conversion candidates

3.4 Features of Niboshi

Niboshi has four main features that enable users to input Japanese efficiently on slate devices. We prioritized experienced users as the target users of the system, so the features are aimed towards them rather than novice users. However, since the system is designed to be simple and easy to learn, novice users can quickly learn how to use it.

The four main features of Niboshi are that it has multi-touch input, enables the device to be firmly held with both hands while text is input, can be used without visual confirmation of the input buttons, and has a large text display area with a small interface.

- **Multi-touch input**

Multi-touch input in Niboshi enables experienced users to input a character as quickly as by one stroke of one finger. The order in which the consonant and vowel are touched does not matter because the system has robustness for the touch order.

- **Device can be firmly held with both hands**

Niboshi is designed so that users can still hold the device with both hands while using Niboshi. All of the input operations can be done by two thumbs, without change to how the device is held.

- **Visual confirmation of buttons unnecessary**

Niboshi has a calibration function for aligning the input buttons to the locus of where the user's thumbs touch. The system also uses visual feedback in the text display field to show which buttons the user is holding, so characters can be input without visual confirmation of the buttons being pressed. Both the calibrated button layout and the visual feedback in the text display field have the effect to reduce the number of Focus Of Attention (FOA) [9]. QWERTY requires users to view both the text display field and the software keyboard. In contrast, Niboshi requires users to only view the text display field; this can reduce the stress and improve the speed of inputting text.

- **Large text input display area**

The buttons in Niboshi are separated into two sets, one at either edge of the screen. This leaves the middle of the screen open for the text display area, unlike QWERTY, which takes up a full rectangle of space on the screen.

3.5 Calibration

There are more ways of holding slate devices than smartphones or PDAs because slate devices are bigger, thus offering more places for the hands to grasp. They are also heavier, so how they are held depends on the size and the shape of the user's hands.

Therefore, we found that the button layout should be able to be calibrated for each user. We implemented a calibration function in Niboshi by which the system can adjust its button layout for each user depending on the kinematic characteristics of human thumbs. This enables users to hold the device as they want, and thus the system is flexible as to how and where it is held.

To calibrate the system, users hold the device in both hands as they want and are then directed to move their thumbs on the screen (Fig. 7a). Then, the system calculates a button layout from the paths of the thumbs (Fig. 7b).

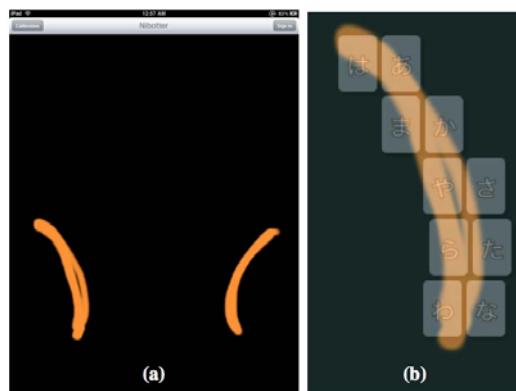


Fig. 7. User's thumb path on screen (a) and calibrated button layout (b)

4 Implementation

We have implemented Niboshi in Objective-C and iOS SDK 4.2. The application runs as Twitter client on iPad. The application consists of the interface, hiragana/katakana input engine, kanji conversion, button calibration and twitter engine. Sticky hit check function is implemented to stabilize the hit check from two hands. We found that a user often input a character that he/she didn't intend in a pilot study. This is because a point touched by a thumb moves small amount when the other thumb releasing. To avoid this problem, we implemented sticky hit check. Sticky hit check is the function that enlarges the hit check area of a button when the button is touched (Fig. 8).



Fig. 8. Sticky hit check

5 Evaluation

We conducted an experiment with one user to evaluate the learning curve of Niboshi. The subject is one of the authors who is used to computer operation. The task of the experiment is text-copying. The time to copy an article (206 characters long) from a news web site using Niboshi, and he is instructed to do at least one session every day,. The experiment was conducted for 56 sessions for 56 days.

5.1 Results of the Experiment

Fig. 9 shows the result of the experiment, and indicates the speed of input using Niboshi. The horizontal axis shows the sessions, and the vertical axis shows the number of characters input per minute (cpm).

We estimate the learning curve of this experiment from power trendline of the result that is shown as bold line in Fig. 9. The learning curve shows that our system can be fairly learnt in continuous usage.

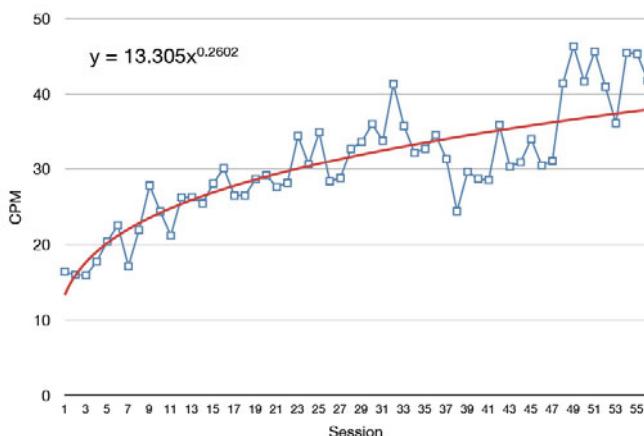


Fig. 9. Input speed of Niboshi

The maximum input speed through the experiment was 46.3 cpm. The input speed increases as the session goes, so we believe that users can learn how to input in Niboshi quickly enough.

6 Conclusion

We developed Niboshi, a Japanese input method for slate devices. Niboshi has four features that improve the overall performance of Japanese input on slate devices. We also conducted an experiment and made sure the learning rate of the system and it indicates that the system can be learnt fast enough to be used on daily basis.

References

1. Kimioka, G., Shizuki, B., Tanaka, J.: Japanese input method using multi-touch for mobile devices and its evaluation. In: IPSJ SIG Technical Report vol. 2010-HCI-138 pp.1–6 (2010)
2. Blickenstorfer, C.H.: Graffiti: In Wow! Pen Computing Magazine, pp. 30–31 (1995)
3. Zhai, S., Hunter, M., Smith, B.A.: The metropolis keyboard - an exploration of quantitative techniques for virtual keyboard design. In: Proceedings of the 13th Annual ACM Symposium on User Interface Software and Technology, UIST 2000, New York, pp. 119–128 (2000)
4. Shiohara, Y., Tano, S., Ichino, J., Hashiyama, T.: Eyes-free one stroke thumb text entry based on the movement characteristic. In: Proceedings of Human Interface Symposium 2007, pp. 573–576 (2007)
5. Tanaka, M., Tagawa, K., Yamanaka, S.: A study on developing of new thumb keyboard TagType. In: Proceedings of the Japan Society of Mechanical Engineers 2001, pp. 112–116 (2001)
6. Shin, H., Lee, W., Lee, G., Cho, I.: Multi-point touch input method for korean text entry. In: Proceedings of the 27th International Conference Extended Abstracts on Human Factors in Computing Systems, CHI 2009, New York, pp. 3871–3876 (2009)
7. Roudaut, A., Lecolinet, E., Guiard, Y.: Microrolls: expanding touch-screen input vocabulary by distinguishing rolls vs. slides of the thumb. In: Proceedings of the 27th International Conference on Human Factors in Computing Systems, CHI 2009, New York, pp. 927–936 (2009)
8. Masui, T.: A Fast Text Input Method for Pen-based Computers. In: Proceedings of Workshop on Interactive Systems and Software 1997, pp. 51–60 (1997)
9. Mackenzie, I.S., Tanaka-Ishii, K.: Text Entry Systems. Morgan Kaufmann, San Francisco (2007)

An Investigation on Requirements for Co-located Group-Work Using Multitouch-, Pen-Based- and Tangible-Interaction

Karsten Nebe¹, Tobias Müller², and Florian Klompmaker³

^{1,3} University of Paderborn, C-LAB, Fuerstenallee 11, 33102 Paderborn, Germany

² University of Paderborn, Warburger Straße 100, 33098 Paderborn, Germany

{Karsten.Nebe,Florian.Klompmaker}@c-lab.de,

tmueller@uni-paderborn.de

Abstract. Cooperation and coordination is crucial to solve many of our everyday tasks. Even though many computerized tools exist, there is still a lack of effective tools that support co-located group work. There are promising technologies that can add to this, such as tabletop systems, multitouch, tangible and pen-based interaction. There also exist general requirements and principles that aim to support this kind of work. However these requirements are relatively vague and are not focused on concrete usage scenarios. In this study a user centered approach has been applied in order to develop a co-located group work system based on those general requirements but also on a real use case. The requirements are transformed into concepts and a running prototype that was evaluated with users. As a result not only the usability of the system has been proven but also a catalogue of even more specific requirements for co-located group work systems could be derived.

Keywords: multitouch, tangible interaction, collaboration, cooperation, co-located group work, requirements, user centered design.

1 Introduction

Even though many computerized communication tools exist, there is still a lack of effective tools that support co-located group work. In many situations people fall back on traditional tools such as paper and pen as well as white boards, etc. The strength of these traditional tools is certainly the usability (as the use of the tools is obvious to everyone) as well as the support of workspace awareness (almost everyone can see and understand what the other group members do). Usual PCs are designed for single-user-use and do not support collaborative and social aspects – known as co-located groupware – in an adequate way.

There are new types of computer systems that might add to this. Multitouch technology makes it possible to simultaneously perform multiple actions through a touch-based interface. This offers new ways of providing input that are not present in classical devices such as the keyboard and the mouse. For instance, it allows users to manipulate screen objects using several fingers at once. There is another approach,

called tangible interaction that enables users to interact with digital information through the physical environment. This is often realized by using physical objects that can be placed and/or moved on a surface. The physical interaction is recognized by the system and results in a manipulation of virtual objects of the graphical user interface. Especially in co-located groupware this new interaction techniques can become a promising addendum as it offers a special quality of awareness – users can see who is going to manipulate information directly. In addition, there is also pen-based interaction which is (still) common on many mobile devices or even tablet-PC's that might add to scenarios of computerized co-located groupware.

In general, the authors believe that multitouch-tables are useful in situations where a group of people physically comes together in order to cooperatively work on a mutual data base while aspects of social interaction is essential to achieve a common goal. These ‘new’ technologies seem to be promising for co-located group work systems. However, there is a lack of realistic use scenarios showing how to exploit the advantages of these new technologies and its ways of interaction. The aim of this study is to transfer the requirements gathered from user analysis and literature research into concepts and a running prototype in order to testify the solution with real users. Beside this, a catalogue of concrete requirements for the development of co-located group work systems will be presented.

2 Existing Work on Collaborative Group Work

During the last years substantial research on tabletop groupware has been performed providing knowledge about the design and usage of tabletop systems (e.g. see [1]). Many studies are technology focused and evaluate whether and how certain technologies can be used at all. An example for this are studies that deal with ‘privacy protection’, such as [2] or [3]. They give a pretty good impression how privacy can be protected but they do not consider when the concepts are useful or not. Other studies focus on intergroup development in terms of system interaction concepts, e.g. ‘release’, ‘relocate’, ‘reorient’, ‘resize’ [4] or ‘comprehension’, ‘coordination’ and ‘communication’ [5]. These studies both provide solid information about human interactions at tabletop systems and prove their results empirically. However, there is a lack of reasonable usage scenarios, which makes it difficult to rate the relevance of results beyond the experimental conditions.

While looking at current studies it could be said that only few studies actually focus on real-life tasks and put them into context. In many cases the kind of interaction, which is tested, is far from current practice and the underlying tasks are basically tailored around the subject of research. As a result, this does not show how technology actually can support human group work in general. In addition to the studies that are more or less technology centered there are pretty good and often-cited studies on general requirements that need to be considered while developing group work. Most of them aim to define basic requirements of group work, which help to understand the basic principles of group work at tabletop systems. Scott et al. [7] define eight general system guidelines for co-located, collaborative work on a tabletop display: ‘Support Interpersonal Interactions’, ‘Support Fluid Transactions

between Activities', 'Support Transitions between Personal and Group Work', 'Support Transitions between Tabletop Collaboration and External Work', 'Support the Use of Physical Objects', 'Provide Shared Access to Physical and Digital Objects', 'Allow flexible User Arrangements', and 'Support Simultaneous User Actions'. Morris adds to these general requirements as she defines three key design challenges, which must be solved by tabletop systems to successfully support group work [8]: 'Integrating Public and Private Information', 'Managing Display Elements', and 'Mediating Group Dynamics'.

In summary, it can be said that there are meaningful general requirements that appear to be a solid basis for the development of co-located collaborative group work systems. However, there are fewer studies that focus on real use cases in order to identify the users needs and transfer them into such systems. However, from a user centered design perspective further requirements will occur while looking on the context of use in real-life situations. Therefore the authors applied interviews and questionnaires in order to transfer the results into a running prototype for evaluation.

3 Proceedings

To develop usable applications that support co-located collaborative group work, it is necessary to analyze the current real-life situation. This is important to understand the context of use and to gather the knowledge needed to design and implement a solution, which fits the users requirements. Therefore a common real-life scenario has been identified, which exists in practice: Today, students very often have to create presentations or reports as a group within the context of university courses. They usually start off with a given topic of their work but they are encouraged to find more deepen information on their own, e.g. through a literature research. This information then must be compiled, combined in new ways, rearranged etc. Usually not all of the work is done in a group. Specific tasks are distributed and performed by members of the group. This not necessarily needs to happen during physical meetings of the group but also separately. However, the results of the work always have a share in the final presentation or the report.

Of course, this scenario is not restricted to students group work. In many other areas of work like research or software development similar situations exists in which a group of persons has to gather information and rework it into a composed result. Thus the authors believe that it will be easy to transfer the experiences to other fields of work. To properly understand the context of use, interviews have been performed with a group of students working on collaborative tasks in a non-IT environment.

3.1 Interviews

Based on the group work scenario of creating a presentation or report interviews have been performed to gather more information on how the group work is done, what is done in group work and what is done alone, which media is used for communication, how the gathered information is assorted and what tools are used.

Six students were interviewed about their experience with group work during their studies. The interview was separated in two parts: The focus of the first part of the interview laid on the organizational part of the group work, like how often they meet, which work is done within the group or which work was done alone. This helped to get a detailed view on how a tabletop system can potentially support the process of group work and which types of tasks are performed during this time. The second part was about the media and artifacts the students use to communicate, especially in the times between their meetings and in which way they organize their information.

As a result it turned out that organizing information, reconciling work results and discussions about results as well as decisions about further actions are generally done as part of the group work. Creating texts, graphics or slides or doing (long-winded) research is done more often individually. This leads to the conclusion that a tabletop system for co-located collaborative group work especially needs to support working with previously created information and less with the creation of many new artifacts. This also means that it must be easily possible to integrate the artifacts created by the individual group members into the group work process. Between the meetings, the students use asynchronous communication, especially email, more often than chat and telephone to communicate with each other. The reason is, that email not only serves as communication medium. Usually additional artifacts are being attached that represent the individual contributions to the group work. This implies, that individual artifacts should be accessible during the group meetings. The common artifacts the interview partners named are tables, lists or plain papers, which must be supported by the system consequently. As a result of the interviews the most important requirements from the use case perspective are as follows: “Provide the ability to exchange objects between users”, “Specialized types of artifacts must be supported”, “Support access to data from asynchronous communication”, “Support merging and distributing of work” and “Favor functions for working with existing artifacts over functions for creating new artifacts”.

The knowledge about the scenario and the details from the interviews result in user requirements that have major impact on the concepts for the final system, which are being described in the next paragraph.

3.2 Concepts and Prototype

The requirements derived from the interviews have been the starting point for general concept ideas of the future system. In addition to the user requirements, general requirements for the development of co-located collaborative group work systems have been taken into account (see section 2). While looking on these requirements and keeping current interaction technologies into account the concepts for the final tabletop groupware system have been developed. As the concepts are very extensive, only a short example is given.

As a result of the interviews it was known, that the users often use plain sheets of paper within the context of the group work. As the system is intended to support the already existing group work process, it needs to reflect this and to provide the users with a paper function, which does not force them to change their typical way of group work. They should be able to draw with pens as they always do. Also paper can be

easily moved to any place, turned in any direction or handed over from one person to another. Consequently these functionalities have been transferred into the concepts. While developing the concept of virtual paper the general requirements have been taken into account. Some of them lead through an extension of the functionality, some were used to backup design decision. For example, concerning the orientation of the paper, it is important to allow the users to freely move and rotate the paper to coordinate their actions, communicate with each other and support the understanding of information [5][9]. This fits perfectly with the results based on the interviews. Other publications add new helpful requirements. Tang et al. [6] found out, that users often switch between more detailed views for individual work and more global views for group work. As a consequence the ability to group and scale the digital paper by simple gestures has been added to the concepts. An additional requirement was to easily be able to switch between different activities during the process, e.g. writing, painting, moving or grouping objects, etc. Thus, we added different types of interaction to give users more flexibility, for example the ability to draw or write using a pen or the fingers (depending on the task and the needed precision). Finally the possibility to create copies of a virtual paper has been added in order to offer more individual access points to the system [10, 11, 12] and to allow the users to move from the closely to more loosely coupled group work [7]. More details can be found in Müller, 2010 [13].

Further concepts, e.g. for a ‘table-sheet’ and ‘check-list’ as specialized types of artifacts as well as generic functions such as a home- and storage-areas have been developed. These concepts all have one in common: They base on the users’ needs but add additional necessary functionality based on existing research (compare section 2). The final concepts then were transferred into a running prototype. A self-made interactive table was used that consists of a 55” full HD screen that supports multitouch, digital pens and tangible objects as input for interaction (www.useTable.de). The software was developed in C#. To give the reader an impression, some pictures of the concepts and the final prototype on the useTable are shown in Figure 1 and Figure 2.



Fig. 1. and **Fig. 2.** Drafts of virtual paper, list within a storage area; Prototype running (right)

3.3 Evaluation and Results

In order to evaluate the developed concepts, usability tests with the prototype were conducted. As stated, the goal of this study was to support the group work of students

through the use of a tabletop system. The usability test was limited to one group meeting (as the usual time frame for the groups to solve their task stretches over weeks). Each group had the predefined starting point: A group of students has the task to give a talk concerning software and ergonomics; As a part of their presentation, they want to include which kinds of icons one can distinguish. Therefore they want to show some examples of the three different types of icons ‘semi-abstract’, ‘abstract’ and ‘symbolic’ (e.g. icons for load, save, edit and delete). The participants were provided with some pre-work results they could access during the test to simulate their individual work. This included a classification of different representations for icons. The goal of the group meeting was to create draft versions of the four icons for each of the three types. The task and the duration period (25 mins) were set to offer enough flexibility studying the test persons using the system. There was no specification how the users should proceed to solve the task.

Twelve tests with each two test-persons were conducted, whereas all users studied computer science. At first each group was introduced to the system and got time to explore the functions. After that, they were given their task to accomplish. During the test, the users were observed and notes were taken. The goal was to identify how and which parts of the system were used, and how frequently the participants were communicating and cooperating. After the test, the participants were asked to fill out a questionnaire regarding their personal perception of the systems features and its appropriateness and usability.

The observation of the tests showed that the different groups had very different approaches towards the usage of the systems. Nearly all the implemented features were used but each group had their own combination. Some for example used the background to draw the icons on while others used the virtual paper. The ones who used the paper sometimes used the storage-area-function to organize the artifacts; others just placed them on the center area of the table. In general there occurred very few problems with the usage of the system and most of the discussion was about solving the task. In most cases the icons were drawn cooperatively even though two groups preferred to split the task and work individually. The features provided were considered to be appropriate and easy to use with minor problems concerning text input, which had technical reasons. More details on the test can be found in Müller, 2010 [13].

In general it can be said that the system was accepted by the users and is easily utilized to fulfill the given task. Probably the most interesting result the authors can draw from investigation is that a real-life scenario was found, which could successfully be supported by a specialized tabletop system. This system was created based on user analysis. The resulting user requirements were supplemented with general requirements based on literature and have finally been balanced with technical feasibilities in order to create the overall system. The general requirements turned out to be very useful but are formulated on a very abstract level. To add to this, the authors created an overview of useful requirements for collaborative and co-active group work and append some new ones based on the experience of this study.

Table 1. Overview of concrete requirements for co-active group work systems, mapped to general requirements from Scott et al. [7] and Morris [8]

															Short-Description	
I/W	Focus (Workspace, Artifact, Input)	Interpersonal interactions	Fluid transactions between activities	Transitions betw. pers. and group work	Use of physical objects	Shared Access to physical and digital objects	Flexible user arrangements	Simultaneous user actions	Integrating public and private information	Managing display elements	Moderating group dynamics	Provide specialized type of artifacts	Input workspace awareness & ergonomic requirement	Requirement	References	
W	Allow personal as well as group work													Users sometimes switch between personal work and group work. Therefore, the system must be able to support closely coupled work as well as very loosely coupled work.	6	
W X X	Provide global and detailed perspectives													Closely coupled groups prefer to work on a more global view while single persons prefer a more detailed view.	6	
I X														Every user must have the ability to contribute to the common work and have access to input devices and/or direct input. There must be enough access points to prevent all of them being blocked at one time.	10, 11, 12	
I X X	Allow parallel input													To enable group members to work loosely coupled, parallel and independent input mechanisms must be provided.	10	
I X	Allow equal access to control for all group members													If there are privileged users in terms of access to control the quality of the group work decreases. Thus all users must be handled equally by their access to control.	10, 14	
W X	Allow equal and open access to information for all group members													If there are users with privileged access to information, the group will tend to ignore that information in their decision-making because of social aspects.	10, 14	
A X	Conflicts should be handled by software strategies implemented in software	X												Whenever conflicts arise, the tabletop-system should solve them automatically in such a way that no human intervention is need.	8, 15, 16	
W X	Conflicts should not be handled by software	X												Conflicts indicate that a closer coordination is needed. Thus it could be harmful for the group process to handle them.	17, 6	
I	X	Provide alternative input mechanisms (direct & indirect) depending on the screen size												Direct input may not always be the best solution, especially while using large screens/tables. Consequently indirect input can be more efficient.	15, 16, 18, 19	
I	X	Clearly indicate who is modifying (especially when using indirect input)												If the group members cannot easily draw connection between an action and the corresponding actor, the workspace awareness will decrease.	15, 16	
W X	Provide sufficient workspace size allowing the division of private, group and storage territories													In group work situations humans tend to divide their workspace into their personal territories for personal things and personal work, group wide territories for common work and storage territories to store currently not used objects.	20	
W X	Provide visible access to all (private and non-private) territories													Every user must be able to openly view all the other territories because otherwise the group process would be hindered.	20	
I X	X	Attach controls as close as possible to the corresponding territories												Set controls and general the functionality for different territories at the very same place, e.g. put the button for clearing the workspace in the group territory as this affects everybody.	20	
A X X X	Support the ability to continuously rotate objects													Humans rotate objects for several reasons. The behavior of the physical objects must be transferred into software.	5, 9	
A X	Provide the ability to exchange objects betw. users	X												Users must be able to transfer objects between each other.	13	
A		X												When objects are exchanged between users, one could use the techniques release, relocate, reorient and resize. Relocate proved to be fastest and least error-prone. Users must be identifiable.	4	
A X	Use transparency to avoid interference	X												When objects are moved, resized etc. they should become partly transparent to minimize interference with other users.	21	
A	Objects must be maintained persistent over multiple user sessions													Sometimes users are going to have multiple sessions about the same topic. This makes it necessary to store the state of the workspace between the sessions.	22	
W X X	Workspace should be totally visible whereas privacy should only be supported, if the group process needed													A transparent workspace is needed to keep everybody in the group at the same level of information, thus privacy can hinder needed	20, 23	
I X	Use tangibles as an access point for interaction as they support group work.													Humans are used to work with physical objects, therefore tangibles are a way to support humans in working with a computer and reduce their mental effort of using it.	7	
A	Specialized types of artifacts must be supported	X												For task completion humans use specialized types of representation. Adapted types of artifacts should reflect this.	22, 24, 13	

Table 1. (*Continued*)

A	X				Support access to data from asynchronous communication	Between the co-located work sessions, the group members will likely communicate via services like email. Provide access to this data during the sessions to harness this data.	13
A	X				Support merging and distributing of work	Meetings are often used, to merge work done by different person or to distribute the tasks among the group members. This means, the system must be able to handle merging work results and distributing tasks + related information.	13
I		X	X		Tangibles are a possible embodiment for a user	Tangibles can be used to easily represent the position of the users on the Table. Each user has a tangible and where ever the tangible is set, the users position is assumed.	13
W	X		X		Tangibles are a way integrate the private data of a user	Tangibles can be used to integrate the private data of a user at the point where the tangible is set on the tabletop computer.	13
A	X			X	Prefer functions for working with existing artifacts over functions for creating new artifacts	Users will likely create the texts, graphics etc. in individual work and during group work they will use this artifacts for discussions or reconcile results.	13

4 Requirements for Collaborative and Co-active Group Work

As part of the research for this study the authors did a lot of requirement engineering to find out how the concepts needs to be designed. The main question was, how to support the users (the students) successfully during their group work (here: creating a presentation or a report). Two sources of requirements were used to define the solution: first, user requirements based on interviews; and second, general requirements based on appropriate publications (compare section 2).

Based on the authors' experiences and as the result of the process and the evaluation it can be said that those general requirements are valid even if (or because of) they are not very strictly formulated. To some kind of degree they offer leeway for interpretation. They worked quite well for the given scenario in this study. However, this study can add to this. Based on the use case those general requirements can be even more substantiated in order to serve as a basis for similar systems to be developed (by other authors). In addition to the general requirements of Scott et al. and Morris two new ones have been created: 'Provide specialized types of artefacts based on common use cases' and 'Input must support workspace awareness and ergonomic needs'.

Furthermore, a supplemental categorization of the requirements has been conducted addressing the typical system-parts, which are: 'Workspace', 'Artifact', and 'Input'. This helps to focus and to distinguish depending on the subject of and context for implementation. In order to prove these even more concrete requirements not only from the practical but also from the theoretical perspective, confirmation in literature has been searched and found. All this information has been consolidated in a condensed overview, which is shown in Table 1. By using this overview, designers and developers will be guided in terms of selecting and translating appropriate requirements for systems that support collaborative and co-active group work. The mapping with general requirements and the categorization therewith help to use this from different perspectives, the theoretical and the practical ones.

Of course, this is not a complete list of requirements but it is a solid base, to be adopted and extended depending on the use case and types of interaction.

5 Summary and Outlook

In this study a user centered approach has been applied in order to develop a co-located group work system. It bases on general requirements derived from literature

research and from task analysis with real users, likewise. A prototype has been developed in order to evaluate: a) the systems usability and b) whether the requirements (especially those based on literature) are fulfilled by the solution.

As a result the requirements were generally fulfilled and the system's features provided were considered to be appropriate and usable. While supporting a real-life scenario the system was accepted by the users and is easily utilized to fulfill the given task. Additionally the authors created a list of requirements for co-located based on the experience of this study. This overview helps designers and developers in selecting and translating appropriate requirements for such systems.

In future the authors will extend the systems functionality based on future user analysis. New concepts for asynchronous and synchronous use cases will be added and the connectivity to web-shared-folders will be implemented in order to provide access to any users' private artifacts. And of course further concrete requirements for co-active group work systems will be developed and evaluated.

References

1. Müller-Tomfelde, C. (ed.): *Tabletops - Horizontal Interactive Displays*. Springer, Heidelberg (2010)
2. Smith, R.T., Piekarski, W.: Public and private workspaces on tabletop displays. In: *Proceedings of the Ninth Conference on Australasian User Interface (AUIC 2008)*, vol. 76, pp. 51–54. Australian Computer Society, Inc., Darlinghurst (2008)
3. Shoemaker, G.B.D., Inkpen, K.M.: Single display privacyware: augmenting public displays with private information. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2001)*, pp. 522–529. ACM, New York (2001)
4. Ringel, M., Ryall, K., Shen, C., Forlines, C., Vernier, F.: Release, relocate, reorient, resize: fluid techniques for document sharing on multi-user interactive tables. In: *Extended Abstracts on Human Factors in Computing Systems (CHI 2004)*, pp. 1441–1444. ACM, New York (2004)
5. Kruger, R., Carpendale, S., Scott, S.D., Greenberg, S.: Roles of Orientation in Tabletop Collaboration: Comprehension, Coordination and Communication. *Comput. Supported Coop. Work* 13(5–6), 501–537 (2004)
6. Tang, A., Tory, M., Po, B., Neumann, P., Carpendale, S.: Collaborative coupling over tabletop displays. In: Grinter, R., Rodden, T., Aoki, P., Cutrell, E., Jeffries, R., Olson, G. (eds.) *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2006)*, pp. 1181–1190. ACM, New York (2006)
7. Scott, S.D., Grant, K.D., Mandryk, R.L.: System guidelines for co-located, collaborative work on a tabletop display. In: Kuutti, K., Karsten, E.H., Fitzpatrick, G., Dourish, P., Schmidt, K. (eds.) *Proceedings of the Eighth Conference on European Conference on Computer Supported Cooperative Work (ECSCW 2003)*, pp. 159–178. Kluwer Academic Publishers, Norwell (2003)
8. Morris, M.J.: Supporting Effective Interaction with Tabletop Groupware. Ph.D. Dissertation. Stanford University, Stanford (2006)
9. Kruger, R., Carpendale, S., Scott, S.D., Greenberg, S.: How people use orientation on tables: comprehension, coordination and communication. In: *Proceedings of the International ACM SIGGROUP Conference on Supporting Group Work (GROUP 2003)*, pp. 369–378. ACM, New York (2003)

10. Hornecker, E.: A design theme for tangible interaction: embodied facilitation. In: Gellersen, H., Schmidt, K., Beaudouin-Lafon, M., Mackay, W. (eds.) Proceedings of the Ninth Conference on European Conference on Computer Supported Cooperative Work (ECSCW 2005), pp. 23–43. Springer, New York (2005)
11. Inkpen, K.M., Hancock, M.S., Mandryk, R.L., Scott, S.D.: Collaboration Around a Tabletop Display: Supporting Interpersonal Interactions. In: Simon Fraser University Tech. Report (2001)
12. Marshall, P., Rogers, Y., Hornecker, E.: Are Tangible Interfaces Really Any Better Than Other Kinds of Interfaces? In: Proceedings of the the Workshop on Tangible User Interfaces in Context and Theory at CHI 2007 (2007)
13. Müller, T.: Untersuchung zur kolokalen, kollaborativen Wissensarbeit am Gegenstand von interaktiven Multi-Touch- und Multi-User-Tables, Master's thesis, University of Paderborn (2010)
14. Rogers, Y., Hazlewood, W., Blevis, E., Lim, Y.-K.: Finger talk: collaborative decision-making using talk and fingertip interaction around a tabletop display. In: CHI 2004 Extended Abstracts on Human Factors in Computing Systems (CHI 2004), pp. 1271–1274. ACM, New York (2004)
15. Nacenta, M.A., Pinelle, D., Stuckel, D., Gutwin, C.: The effects of interaction technique on coordination in tabletop groupware. In: Proceedings of Graphics Interface 2007 (GI 2007), pp. 191–198. ACM, New York (2007)
16. Pinelle, D., Nacenta, M., Gutwin, C., Stach, T.: The effects of co-present embodiments on awareness and collaboration in tabletop groupware. In: Proceedings of Graphics Interface 2008 (GI 2008), pp. 1–8. Canadian Information Processing Society, Toronto (2008)
17. Hornecker, E., Marshall, P., Sheep Dalton, N., Rogers, Y.: Collaboration and interference: awareness with mice or touch input. In: Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work (CSCW 2008), pp. 167–176. ACM, New York (2008)
18. Ha, V., Inkpen, K.M., Whalen, T., Mandryk, R.L.: Direct Intentions: The Effects of Input Devices on Collaboration around a Tabletop Display. In: Proceedings of the First IEEE International Workshop on Horizontal Interactive Human-Computer Systems (TABLETOP 2006), pp. 177–184. IEEE Computer Society, Washington (2006)
19. Toney, A., Thomas, B.H.: Applying reach in direct manipulation user interfaces. In: Kjeldskov, J., Paay, J. (eds.) Proceedings of the 18th Australia conference on Computer-Human Interaction: Design: Activities, Artefacts and Environments (OZCHI 2006), pp. 393–396. ACM, New York (2006)
20. Scott, S.D., Sheelagh, M., Carpendale, T., Inkpen, K.M.: Territoriality in collaborative tabletop workspaces. In: Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work (CSCW 2004), pp. 294–303. ACM, New York (2004)
21. Zanella, A., Greenberg, S.: Reducing interference in single display groupware through transparency. In: Prinz, W., Jarke, M., Rogers, Y., Schmidt, K., Wulf, V. (eds.) Proceedings of the Seventh Conference on European Conference on Computer Supported Cooperative Work (ECSCW 2001), pp. 339–358. Kluwer Academic Publishers, Norwell (2001)
22. Leonardi, C., Pianesi, F., Tomasini, D., Zancanaro, M.: The Collaborative Workspace: A Co-located Tabletop Device to Support Meetings. In: Waibel, A., Stiefelhagen, R. (eds.) Computers in the Human Interaction Loop, pp. 187–205. Springer, London (2009)
23. Gutwin, C., Greenberg, S.: A Descriptive Framework of Workspace Awareness for Real-Time Groupware. Comput. Supported Coop. Work 11, 411–446 (2002)
24. Hampel, T., Niehus, D., Sprotte, R.: An Eclipse based Rich Client application for open-sTeam and its real world usage. In: Pearson, E., Bohman, P. (eds.) Proceedings of the World Conference on Educational Multimedia, Hypermedia & Telecommunications, pp. 1304–1309. ED-MEDIA, Chesapeake (2006)

Exploiting New Interaction Techniques for Disaster Control Management Using Multitouch-, Tangible- and Pen-Based-Interaction

Karsten Nebe¹, Florian Klompmaker¹, Helge Jung², and Holger Fischer¹

¹ University of Paderborn, C-LAB, Fürstenallee 11, 33102 Paderborn, Germany
`{karsten.nebe,florian.klompmaker,holger.fischer}@c-lab.de`

² SIS GmbH, C-LAB, Fürstenallee 11, 33102 Paderborn, Germany
`helge.jung@c-lab.de`

Abstract. This paper shows the proceedings and results of an user centered design process that has been applied in order to analyze how processes of management in disaster control can be optimized while using new interaction techniques like multitouch, tangible and pen-based interaction. The study took part in cooperation with the German Federal Agency for Technical Relief. Its statutory tasks include the provision of technical assistance at home and humanitarian aid abroad. Major focus of this work is the IT-support for coordination and management tasks. As result we introduce our prototype application, the software- and hardware requirements towards it as well as the interaction design that was influenced by the outcome of the user centered design process.

Keywords: Interaction techniques, multitouch, tangible interaction, pen interaction, disaster control management, THW, user centered design.

1 Introduction

The development of innovative products is often driven by new technology while loosing the focus on identifying reasonable use cases from the early stage. However, this is significant in order to persist on the market. Real use cases are necessary to identify the potential of the new products and to determine whether they are useful and support the users in an efficient and effective manner. Based on the use cases technical details can be evaluated and thus, appropriate interaction techniques identified. In comparison to standard WIMP (Windows, Icons, Menus and Pointer) based interaction, there exist various ‘new’ techniques for the users, such as multitouch, tangible and pen-based interaction.

Multitouch technology makes it possible to perform multiple interactions through a touch-based interface simultaneously. Since input and output are co-located it offers a direct interaction with the visualized data. Hence interaction techniques can be designed to be very intuitive and natural. When Apple’s iPhone appeared on the market in 2007, multitouch technology became very fashionable even though the technology exists since 1982 [1]. Nowadays devices with different form factors, from smart phones through tabletops to large interactive walls are omnipresent. Some

devices, especially tabletops and interactive walls, allow multiple users to work at one system at the same time. Many tabletops allow also tangible interaction by placing physical objects on the screen that serve as input devices [2]. Such objects can adopt various form factors and thus, can also be designed in the style of physical tools users are familiar with. They provide tactile feedback and can overcome the drawback of occlusions caused by the finger that occur on classical touchscreens [3,4,5].

To a certain degree digital pens are special tangible interaction devices. Most humans are familiar with using pens in their every day life. The benefits of using pen and paper are manifold: natural, fast, precise and usable. Therefore digital pens are used for many tasks that require high precision or text input, too. Nowadays, even large touch displays can be equipped with digital pens for precise input, e.g. [6], and therewith enable the users to interact and manipulate directly.

Nevertheless, as mentioned before there is a lack of realistic use cases describing how to exploit the advantages of systems that commonly use such a variety of new interaction techniques. While looking on available ‘demonstrators and commercial applications’, most have ported existing IT-applications to the new technology without exploiting its true advantages. As a result the majority of such systems still rely on the ‘wow factor’, failing to show the benefits and improved efficiency of these new interaction techniques. To verify these techniques we build a hardware device and investigated in real scenarios of use.

1.1 The useTable – A Multi-user Tabletop Supporting Multitouch, Tangible and Pen-Based Interaction

We have built our own multitouch-table - called the ‘useTable’¹ - and we are investigating scenarios, which can generate an increased efficiency for the users. In general, we believe that the useTable is useful in situations where a group of people physically comes together in order to cooperatively work on a mutual data base while aspects of social interaction between the participants is essential to achieve a common goal.

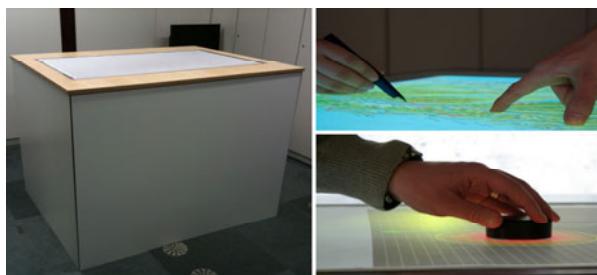


Fig. 1. The useTable supports multitouch, tangible and pen-based interaction

The useTable (see Figure 1) consists of a 55” display that offers full HD image projection. The projector is mounted beneath the surface and the image is projected to the top using two mirrors. For finger-tracking FTIR (Frustrated Total Internal Reflection)

¹ <http://www.usetable.de>

is applied and objects on the surface are tracked using a combined DI (Diffused Illumination). The camera on the bottom of the table is equipped with a corresponding IR filter and is connected to a tracking PC that applies intelligent filter routines. The projection surface is equipped with an antireflex diffusor sheet that enables pen-based interaction by using Anoto² digital pens [7].

The useTable enables us to investigate and evaluate in various interaction techniques and to prove the usability in different scenarios.

1.2 Finding Meaningful Use Cases

We generally believe that the added value of the useTable can be found in scenarios in which various people physically comes together to achieve a common goal. Scenarios in which coordination and planning are core tasks seemed to be promising, and consequently end users such as fire departments, police, etc. occurs to our mind. However it turned out that many processes and workflows in industry and public authorities are highly automated and standardized as they often based on strict rules and regulations. Thus, the idea of sharing a common database and discussing details in a group in order to plan and prepare next steps does not always match for most incidents. However, there are extraordinary situations in which those standard workflows cannot be applied, such as disasters like flooding, large-scale fires, etc. In cooperation with the German Federal Agency for Technical Relief (German abbr. 'THW') we investigated in such unpredictable scenarios. The THW's statutory tasks include the provision of technical assistance at home and humanitarian aid abroad. Their work during an operation requires planning, coordination, collaboration and coactivity. Today they predominantly make use of classical tools like pen and paper, paper maps, boards and magnetic labels (see Figure 2). Most of the dedicated workflows and roles are also standardized hence IT tools have to be integrated with care since they are not allowed to violate existing specifications.

After finding a project partner and a meaningful use case the goal of the research presented in this paper was to develop a prototype software and hardware solution that fits to the needs of the future users.



Fig. 2. Classical work of the THW: Using paper, pens, paper maps and magnetic labels

² <http://www.anoto.com>

2 Related Work

This section introduces the most important related work in the field of multitouch, tangible interaction, pen-based interaction and disaster control management.

2.1 Multitouch, Tangible and Pen-Based Interaction

Much research has already been spent on interaction design and techniques in the field of multitouch, tangible and pen-based interaction with large screens like tabletops or interactive walls. For example, well-known single- and multitouch techniques for the translation and rotation of virtual objects have been developed and evaluated [8,9,10]. These techniques are already state-of-the art in commercial products and are therefore known by many people so that they can be used without much learning effort for the users. Thus, they can be described as being intuitive and easy. Further on it has been proved that multitouch interaction in multi-user scenarios facilitates awareness [11] because it is easier for everyone to notice *who* (group awareness) is currently doing *what* (interaction awareness).

The awareness can even be improved by using tangible interaction [12] because the interaction input and output is bounded to a specified physical object. Since physical objects can easily be placed on digital tabletops for interaction, numerous example applications exist in this field, from simple board game adaptations to complex multi-user systems supporting collaborative decision-making. These so called tangible user interfaces give physical form to digital information and they build a physical representation and control in one device [13]. Nowadays tangible user interfaces are well explored. It has been found that they are much more natural and useable than user interfaces that are controlled by mouse and keyboard [14] and even design recommendations have been worked out [15, 16] for general purposes.

To a certain degree pen-based interaction is a special kind of tangible interaction. Since we are used to interact with pens on paper, digital pens are a powerful tool for intuitive and precise input on interactive screens. Text input can be realized using digital pens in a very natural way. User centered studies have shown how pen based interaction can be used on interactive tabletops and how intuitive interaction techniques can be investigated [17].

Since interaction requires an adequate representation of information in order to fulfill the user's needs, the visualization has to be designed carefully. Large screens enable multiple people to get the same view on data. Using interaction techniques like the ones previously described also offer multi-user input and hence support co-located collaborative work. It has been shown that different views onto the same data can be helpful to understand complex structures (e.g. [18]). Multiple views can either be realized through software (e.g. multiple windows on the same screen) or through hardware (e.g. the use of multiple screens). Hence carefully designed visualization of information helps to understand complex data. Making information quickly comprehensible is an important task, especially for decision-making during a disaster.

2.2 Disaster Control Management

In the field of disaster control management (DCM) some research studies can be found that address collaboration and coordination using innovative information

technology (IT). Nevertheless only little effort has been spent in designing systems that copy existing real-life-workflows of disaster control organizations like fire departments, police, medical services, etc. The use of new technology is often rare in this area because classical tools like pens, paper, paper-based maps and plastic labels are well proven and failsafe. However the idea of using interactive screens for disaster training purposes already appeared and seems to be a very promising one [19].

An article from the research institute TNO³ introduces some ideas how multitouch tabletops can be an effective assistance for decision-making during a disaster. It has been investigated in which departments such devices could be used, which workflows can be mapped and which not. The study shows that even less IT-experienced people could use the tabletop without much learning effort. The test persons also managed to collaborate successfully using the device. Nevertheless there were also some drawbacks of the system design and we think that these could have been addressed in a much better way by integrating the user into the design process right from the start.

3 User Centered Design Process

User Centered Design (UCD) is an established methodology in the software-industry that focuses on the users of a future system and aims to create solutions that fit the users' needs, their requirements and that support their tasks and goals. According to Jokela [20], the advantages for users are far-reaching and include increased productivity, improved quality of work, and increased user satisfaction. One of the central quality attributes for interactive systems is their usability [21] and the main standardization organizations (IEEE 98, ISO 91) have addressed this parameter for a long time [22]. Especially for tasks that are highly time critical, e.g. DCM, interfaces are required that are easily understandable and intuitively controllable.

In order to create usable solutions it is necessary to involve users in early stages and during the process of development. UCD adds to this by providing different methods applicable at different development stages. Examples include contextual and behavioral analysis (in terms of interviews, site-visits, etc.) in order to gather the users requirements, needs and habits. In addition it is important to know the user's environment and to perform user-tests in order to prove that the solution fits the user's needs.

Together with our project partner THW we performed an UCD approach in order to develop an IT solution that perfectly fits to their needs. We started this process with an initial meeting where we worked out some first application ideas. We got interesting and important insights about the workflows, rules and regulations the THW staff has to follow. Back in our laboratory we created some sketches and functional requirements that we thought would fit the needs and habits of the THW staff. This was followed by a review of THW representatives in order to get the ideas evaluated and to get more detailed information. During a full day practice we then observed the involved people and the workflows in detail. We installed video cameras and microphones in the operations center and communications center and recorded the course of events during the practice. Later on we reviewed the recordings and worked out very detailed workflows of every involved person having a specific role. These

³ [http://weblog.tno.nl/nui/2009/11/10/
how-does-multi-touch-fit-in-crisis-management/](http://weblog.tno.nl/nui/2009/11/10/how-does-multi-touch-fit-in-crisis-management/)

workflows were then transferred in software in terms of our first prototype setup. Here the typical UCD development cycle re-started: The THW representatives reviewed the application running on the useTable and their recommendations were used for the next version of the software and also for minor hardware modifications.

Recently we had a second on-site practice observation at the THW in order to get ideas about further software support tools and refinements. In the following we will explain the outcomes of the UCD process and our prototype application.

4 Results

The division of the THW with which we cooperated within this project mainly deals with water supply and the handling of flooding. They manage the allocation of vehicles, pumps and material. Additionally, they prepare infrastructures, communications and supply. Consequently most working tasks of the THW staff deal with coordination, communication and supply. This section describes the most important findings from the UCD process and it introduces the current prototype application.

4.1 Software Requirements

We found out that multitouch tabletops that enable tangible and pen-based interaction are very suitable for DCM. Regarding the software, many THW-specific tools and workflows can quite easily be supported by software making them more effective. Regarding the hardware, especially the size of the useTable is an important factor. Standing round the table enables a collaborative work and improved communication compared to a vertically installed board.

A very important outcome of our research is that the use of digital maps is very beneficial in this application domain. Digital maps are more flexible than paper maps and they offer comfortable navigation and manipulation. They allow continuous zooming and are infinitely expandable. Most map sources are available via the Internet and also special map sources that have to be officially requested by an emergency organization are available online and thus easily accessible. Finally digital maps allow the visualization of additional layers (e.g. streets, pipelines or weather). To enable the visualization from different map sources and the overlay of several layers we created a framework that prepares map data from different sources to be visualized on the useTable. This framework also enables the creation and display of arbitrary geo-referenced objects like pictures, landmarks or annotations on the map.

Besides the visualization of the map, the THW staff needs to manage and visualize the current situation in the field. Therefore we developed graphical user interfaces that copy their classical way of paper work. We implemented forms, signs and labels that are standardized and familiar to the THW staff. Amongst other things these interfaces can be used to create so called ‘damage accounts’. These show the current situation and the assignment of vehicles and entities to different damages in the field.

Software tools in general can ease complex work in the area of DCM. For example, when dealing with water supply and pumps the THW staff used pens, papers and calculators to estimate the amount of needed tubes and pumps. In the on-site practices we got the impression that this procedure is very time consuming and cumbersome.

Therefore a pump tool was developed that uses geospatial data to calculate the amounts depending on the distance profile (see figure 5). A new hose track can be created with a digital pen directly on the map. The digital pen allows a very precise input and is therefore much more suitable than finger-touch input. This tool was also tested and the THW staff was really stoked about it. This result shows how powerful user centered design can be through observation of potential end users during their every day work. In order to be operated by a user, the different software tools have several requirements towards the underlying hardware.

4.2 Hardware Requirements

As described previously, additional views can enhance the perception of the presented data. During the user observation we found out that the vertical board in the operations center always served as an overview of the current situation to every attendee. Therefore it is crucial to keep this present. We therefore installed an additional beamer in our useTable laboratory (see figure 3) that is connected to the application pc. This second view is not fully equal to the view on the useTable as it only shows the map and the geospatial data like damage accounts. In contrast to the table itself, the projection on the wall shows the current situation as reported from the field and not the situation as planned on the tabletop. All interaction parts like graphical user interfaces are only visible on the table where they are operated.

Regarding the equipment of the useTable we found out that HD image projection is crucial when dealing with texts and detailed digital maps. Further on, network connections for receiving maps and live data from the field are needed. This work is currently in progress in our laboratory. Finally another technical requirement is the use of tangible objects on the useTable. These objects must be highly interactive by allowing the change of their states as described in previous work [23]. How these objects are used for interaction is described in the next section. The technical requirements towards the tracking of fingers and objects can be found in section 1.1.

4.3 Interaction Design

In the previous sections the software and hardware requirements towards the useTable setup is introduced. The most important goal of UCD is the usability of a product. Therefore the interaction capabilities of a system have to be designed carefully. We found some interesting interaction capabilities for the THW use case and the evaluations showed that these are indeed useable and reasonable, too.

Even though multitouch gestures for rotation and translation are well known and accepted we found out that these are not applicable in multi-user map applications. This is due to the fact that multiple fingers on the map would lead to non-comprehensible transformations and to user confusion because actions are not traceable. In our prototype a tangible object (a puck, see figure 4) is placed on the useTable surface. By moving and rotating the puck the map can be translated and zoomed. Since there is only one single puck for map translation, it is always clear to everyone who is currently interacting with the map. This considerably enforces the group and interaction awareness. To enable personal views that allow parallel processing of map tasks by different users we created additional pucks that create a personal window on the screen position where those are placed.

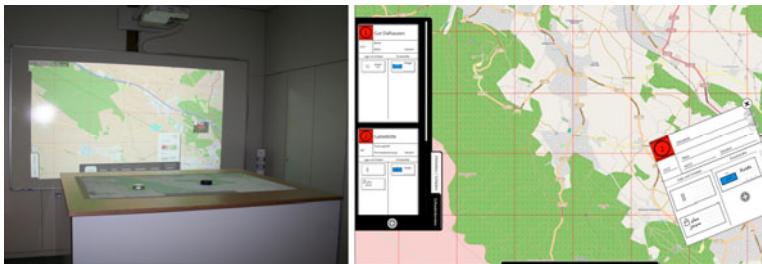


Fig. 3. useTable setup for the THW scenario (left) and appl. showing damage accounts (right)

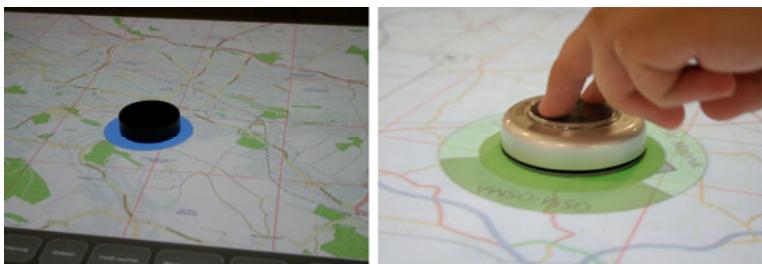


Fig. 4. Map manipulation puck (left) and map source selection (right)



Fig. 5. Editing damage accounts (left) and the pump tool (right)

In order to further explore the possibilities, we developed smart fiducial pucks [23] that possess pushbuttons. These can be used for tasks like selections. Using ring menus and these pucks users can for example change the map source (see figure 4). The awareness of all users is again enforced by the fact that there is only one single map source puck available.

Using classical 2D user interfaces that can be operated with one finger (push buttons, tabs etc.) or multiple fingers (classical translation and rotation) allow users to create and manage objects like damage accounts, entities etc. (see figure 5). During the evaluation phases of the UCD process we found out that most of the THW users have been able to use the interfaces intuitively as they copy existing workflows.

Pen-based interaction is used in our prototype for text entry tasks and therefore in every situation where the users in practice would use a classical pen and paper. Hence

users are not forced to learn new interaction styles. Text input (see figure 5) through digital pens and character recognition works quite well and the information is available as digital text. Further on, pen interaction is used for precise input like map annotations (see figure 5). Since digital annotations can be deleted or edited within a split second they are more practical than classical pen annotations. The representative test users of the system also approved this during the evaluation phases.

5 Conclusion

New and highly interactive technology is always eye-catching and popular through its novel possibilities. However, to survive on the market meaningful use cases and are very important for such kind of hardware. In this paper we describe a user centered design process in which we identified a use case for a multitouch tabletop. Enabling tangible and pen-based interaction, it allows multiple interaction techniques. Together with the German Federal Agency for Technical Relief we developed system that supports disaster control management. We worked out several requirements towards software and hardware and we figured out important aspects regarding the interaction design. The resulting application was reviewed by representatives from the area of disaster control management and is described in this paper in detail.

The next steps in this research project are the use of mobile and in-car devices that are connected to the useTable. Further on we are developing more specific support tools while keeping up the user centered design idea.

References

1. Mehta, N.: A Flexible Machine Interface, M.A.Sc. Thesis, Department of Electrical Engineering, University of Toronto supervised by Professor K.C. Smith (1982)
2. Hornecker, E., Jacob, R.J.K., Hummels, C., Ullmer, B., Schmidt, A., van den Hoven, E., Mazalek, A.: TEI goes on: Tangible and Embedded Interaction. *IEEE Pervasive Computing* 7(2), 91–96 (2008)
3. Wang, F., Cao, X., Ren, X., Irani, P.: Detecting and leveraging finger orientation for interaction with direct-touch surfaces. In: Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology, pp. 23–32. ACM, New York (2009)
4. Brandl, P., Leitner, J., Seifried, T., Haller, M., Dorray, B., To, P.: Occlusion-aware menu design for digital tabletops. In: CHI 2009 Proceedings of the 27th International Conference Extended Abstracts on Human Factors in Computing Systems, pp. 3223–3228. ACM, New York (2009)
5. Vogel, D., Balakrishnan, R.: Occlusion-aware interfaces. In: CHI 2010: Proceedings of the 28th International Conference on Human Factors in Computing Systems, pp. 263–272. ACM, New York (2010)
6. Leitner, J., Powell, J., Brandl, P., Seifried, T., Haller, M., Dorray, B., To, P.: Flux: a tilting multi-touch and pen based surface. In: CHI 2009 Proceedings of the 27th International Conference Extended Abstracts on Human Factors in Computing Systems, pp. 3211–3216. ACM, New York (2009)
7. Haller, M., Brandl, P., Leitner, J., Seifried, T.: Large interactive surfaces based on digital pens. In: 10th International Conference on Humans and Computers, pp. 172–177 (2007)

8. Kruger, R., Carpendale, S., Scott, S.D., Tang, A.: Fluid integration of rotation and translation. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 601–610. ACM, New York (2005)
9. Liu, J., Pinelle, D., Sallam, S., Subramanian, S., Gutwin, C.: TNT: improved rotation and translation on digital tables. In: Proceedings of Graphics Interface 2006, pp. 25–32. Canadian Information Processing Society, Toronto (2006)
10. Reisman, J.L., Davidson, P.L., Han, J.Y.: A screen-space formulation for 2D and 3D direct manipulation. In: Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology, pp. 69–78. ACM, New York (2009)
11. Hornecker, E., Marshall, P., Dalton, N.S., Rogers, Y.: Collaboration and interference: awareness with mice or touch input. In: Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work, pp. 167–176. ACM, New York (2008)
12. Terrenghi, L., Kirk, D., Richter, H., Krämer, S., Hilliges, O., Butz, A.: Physical handles at the interactive surface: Exploring tangibility and its benefits. In: Proceedings of the Working Conference on Advanced Visual Interfaces, pp. 138–145. ACM, New York (2008)
13. Ullmer, B., Ishii, H.: Emerging frameworks for tangible user interfaces. *IBM Systems Journal*, 915–931 (2000)
14. van den Hoven, E., Frens, J., Aliakseyeu, D., Martens, J.B., Overbeeke, K., Peters, P.: Design research & tangible interaction. In: Proceedings of the 1st International Conference on Tangible and Embedded Interaction, pp. 109–115. ACM, New York (2007)
15. Hornecker, E.: A design theme for tangible interaction: embodied facilitation. In: Proceedings of the International Conferences on Computer-supported Cooperative Work, pp. 23–43. Springer, New York (2005)
16. Nebe, K., Müller, T., Klompmaker, F.: An Investigation on Requirements for Co-located Group-Work using Multitouch-, Pen-based- and Tangible-Interaction. In: Proceedings of the HCII 2011. Springer, New York (2011)
17. Frisch, M., Heydekorn, J., Dachselt, R.: Investigating Multi-Touch and Pen Gestures for Diagram Editing on Interactive Surfaces. In: Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces, pp. 167–174. ACM, New York (2009)
18. Roberts, J.: On Encouraging Multiple Views for Visualisation. In: Proceedings of the International Conference on Information Visualisation, pp. 8–14. IEEE Computer Society, Washington (1998)
19. Kobayashi, K., Kakizaki, T., Narita, A., Hirano, M., Kase, I.: Tangible user interface for supporting disaster education. In: Proceedings of SIGGRAPH 2007. ACM, New York (2007)
20. Jokela, T.: An Assessment Approach for User-Centred Design Processes. In: Proceedings of EuroSPI 2001. Limerick Institute of Technology Press, Limerick (2001)
21. Bevan, N.: Quality in Use: Meeting User Needs for Quality. *Journal of System and Software*, 89–96 (1999)
22. Granollers, T., Lorès, J., Perdrix, F.: Usability Engineering Process Model. Integration with Software Engineering. In: Proceedings of the Tenth International Conference on Human-Computer Interaction, pp. 965–969. Lawrence Erlbaum Associates, New Jersey (2002)
23. Klompmaker, F., Nebe, K., Jung, H.: Smart Fiducials: Advanced Tangible Interaction Techniques through Dynamic Visual Patterns. In: Workshop Proceedings of the International Conference on Intelligent User Interfaces 2011, Workshop on Interacting with Smart Objects (2011)

Saving and Restoring Mechanisms for Tangible User Interfaces through Tangible Active Objects

Eckard Riedenklau, Thomas Hermann, and Helge Ritter

Ambient Intelligence / Neuroinformatics Group, CITEC, Bielefeld University,
Universitätsstraße 21-23, 33615 Bielefeld, Germany
`{eriedenk, thermann, helge}@techfak.uni-bielefeld.de`

Abstract. In this paper we present a proof of concept for saving and restoring mechanisms for Tangible User Interfaces (TUIs). We describe our actuated Tangible Active Objects (TAOs) and explain the design which allows equal user access to a dial-based fully tangible actuated menu metaphor. We present a new application extending an existing TUI for interactive sonification of process data with saving and restoring mechanisms and we outline another application proposal for family therapists.

Keywords: Human-Computer Interaction, Tangible User Interfaces, actuated Tangible Objects, Tangible Active Objects, Save and Restore, Menus.

1 Introduction

Nowadays computers and the internet offer a rapidly increasing number of programs and functions with increasing complexity for almost every purpose. Human-Computer Interaction (HCI) becomes more and more important to find understandable interaction metaphors to make these functions more user-friendly. The standard computer workplace consisting of display, keyboard and mouse is not satisfactory on some occasions. Tangible User Interfaces (TUIs) here offer new interaction metaphors by linking digital information to physical objects to make these manipulable by multiple users with their everyday manipulation skills. Since many TUIs use motionless ‘passive’ objects, even basic standard functions we are familiar with from Graphical User Interfaces (GUIs), such as saving and restoring are difficult to implement. In this paper we present actuated Tangible User Interface Objects (TUIOs) that allow implementing these functionalities.

1.2 Table-Top Tangible User Interfaces

Table-top TUI systems in general consist of a surface equipped with some sensing systems for object detection. This can be achieved by visual tracking through a glass surface but there are alternative techniques, such as radio frequency (RF) tracking as used in the AudioPad (1). TUIOs are often simple geometric shapes equipped with markers for tracking. Some systems also allow the visual projection of additional information onto the surface either from above or underneath. Also additional input techniques, such as touch input are also realized in some systems through special

visual markers (2), or multi-touch through Frustrated Total Internal Reflection (FTIR) surfaces, a method originally proposed by Han (3), to name a few. Beyond touch there are systems that also incorporate sound input and output such as the reacTable (2) or AudioPad.

Most table-top TUIs are incapable of saving and restoring arrangements or configurations of TUIOs on the table-top surface. The user either has to remember the arrangement himself, or may take a photo of it and for reconstruction he usually has to reposition the objects manually. To allow the system to save and restore the arrangement of the TUIOs, it needs to be enabled to control them actively. In the following section we will briefly describe existing systems with such actuation possibilities.

2 State of the Art in Actuated TUIs

There already exist systems that cope with the saving and restoring problem or are at least technically able to solve it. PSyBench, the Physically Synchronized Bench from the MIT Media Laboratory is the oldest one (4). Here the positions of objects can be synchronized between two ‘coupled systems’ by a magnetic 2-axis positioning mechanism combined with position sensing realized by a grid of membrane switches. Because of the mechanical design the objects can only be moved one after another.

Another system from MIT Media Laboratory is the Actuated Workbench (5). In this system a special surface is equipped with a grid of individually controllable electromagnets which allow moving ferromagnetic objects around on the surface very quickly. The objects contain a battery and LED for visual tracking from above. Beside classical functions of TUIOs, such as sorting, the authors consider basic functions, e.g. ‘undo’, or teaching and guiding. Furthermore they propose higher level applications, including remote collaboration, simulation and entertainment.

PSyBench and the Actuated Workbench do not allow controlling the orientation of the objects. Weiss et. al. (6) extended the technology used in the Actuated Workbench by using objects equipped with multiple magnets. This allows rotating the objects and controlling properties of the widget like objects, “Midgets”, such as buttons, dials, etc.

The above mentioned projects all use special yet rather low-cost objects. The main disadvantage of this technology is the dependence on a specially constructed and thus expensive surface.

The Planar Manipulator Display (PMD) by Rosenfeld et. al. (7) is another interesting approach for actuating objects. It is technically quite similar to our approach, since they use small-sized mobile robotic platforms. In contrast to our approach, the PMD small-sized mobile robots are slightly bigger than our TAOs and do not represent data themselves, but carry objects they move around. Rosenfeld et. al. also proposed a menu interface to save and restore different arrangements of the objects on the table. The menu used in the PMD system is projected on one side of the interactive surface.

All these systems technically allow saving and restoring arrangements of their actuated TUIOs. Only the PMD offers an implementation for this approach whereas Pangaro et. al. only consider saving and restoring mechanisms for future developments of the Actuated Workbench without offering concrete implementation details.

3 Tangible Active Objects

In this paper we present our Tangible Active Objects (TAOs) (8) and propose a way of tangibly managing multiple object arrangements. Like the objects in the PMD, our TAOs are small-sized mobile robotic TUIOs, which are able to move around on the interactive surface of our tDesk (formerly known as Gesture Desk (9)). The tDesk is equipped with a glass surface and a Firewire camera placed underneath the surface for visual tracking of objects. It is our base platform for table-top TUIs, as shown in Figure 1a).

3.1 Hardware

To build the TAOs' housings we utilized TUImod (10), modular combinable building blocks for TUIOs which were fabricated using a rapid prototyping printer. The electronics are organized in layers with different functionalities. They are physically interconnected through vertical buses which make the design modular and open for future extensions. The main layer is built around the Arduino mini pro¹ platform, a rapid prototyping platform for electronics using an ATmega138 microcontroller. To establish a wireless connection between the TAOs and the host computer another layer was designed carrying an XBee module². The differential drive is controlled by an H-bridge. Between the wheels of the drive a marker is embedded for visual tracking. Fig. 1 depicts the tDesk and the TAOs.

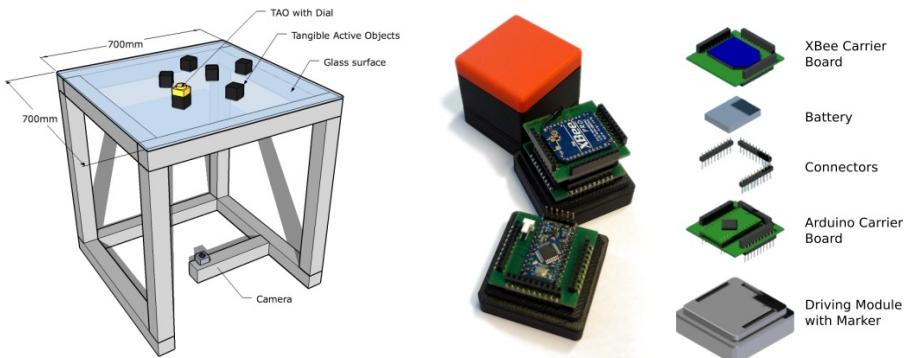


Fig. 1. a) The left most picture depicts the tDesk hardware setup consisting of a cubical table with an edge length of 70cm which is equipped with a glass surface and a Firewire camera for tracking the TAOs from underneath the surface. b) In the center the manufactured devices are shown at different assembly stages (without batteries). c) The right picture shows an explosion drawing of the standard hardware architecture.

3.2 Software Architecture

The TAOs are controlled by a distributed system of software modules, communicating over the XML enabled Communication Framework (XCF) (11), a powerful high-level

¹ <http://www.arduino.cc/en/Main/ArduinoBoardProMini>

² <http://www.digi.com/products/wireless/point-multipoint/xbee-series1-module.jsp>

communication middleware. Fig. 2 depicts an overview over the modules and their communication relations.

The base system is organized in a closed control loop (gray arrows in Fig. 2) starting with the vision module which processes the captured camera images and extracts the marker information (2D position, orientation, and ID) of the TAOs. This marker information is spread over the middleware, connecting the software modules. Other software modules can subscribe to certain information patterns and are notified whenever matching information come in. This content-based notification mechanism allows a loose coupling between the software components. Along with the application modules which are described in the following section, the path planning module subscribes to the visual marker information and computes trajectories and navigation commands which a relay module transmits wirelessly to the TAOs. The trajectories are computed using gradient descent in a potential function whose minimum is located at the target location. Obstacles, such as other TAOs, are automatically avoided by adding repulsive potential hills centered dynamically at obstacle locations. Each step in turn results in a changed camera image and changed marker information and the control loop is closed.

4 Persistent Active User Interfaces

Our first attempt to solve the problem of saving and restoring TAO configurations incorporated an RFID reader on one side of the tDesk and TAOs equipped with RFID tags. The main problem with this approach is that the reader is stationary and not equally accessible by every user. So we changed our implementation and present a new strategy to equip existing TUIs incapable of saving and restoring their configuration. We use two different hardware configurations of TAOs in this approach. The first type of TAOs, depicted in Fig. 1, is the standard configuration consisting of drive, main, and wireless layer. These TAOs are used to interact with the underlying TUI system itself and are referred to as Interaction TAOs (iTAOs) in the following.

To create a fully tangible actuated menu metaphor for persistent tangible representations, we created a new hardware layer with an actuated dial set-top for the TAO which is able to represent different states with an internal degree of freedom and thereby allows the user to select different commands as shown in Figure 3. The dial is implemented using a motorized potentiometer with 300° rotational angle. Therefore it can act simultaneously both as input and an output device and adds an internal degree of freedom to our TAOs. We added basic speech synthesis to convey the menu structure to the user. Moving the menu dial from left to right, a spoken voice feedback reports each menu item. An item is selected by letting the dial rest on the actual item for a few seconds. Again feedback informs the user about the action selection and eventually the dial state changes.

The dial layer is used in the second type of TAOs depicted in Figure 3. These objects are identified via a Fiducial marker, and do not have the differential drive of the iTAOs. So these TAOs cannot move around on the table-top surface and their

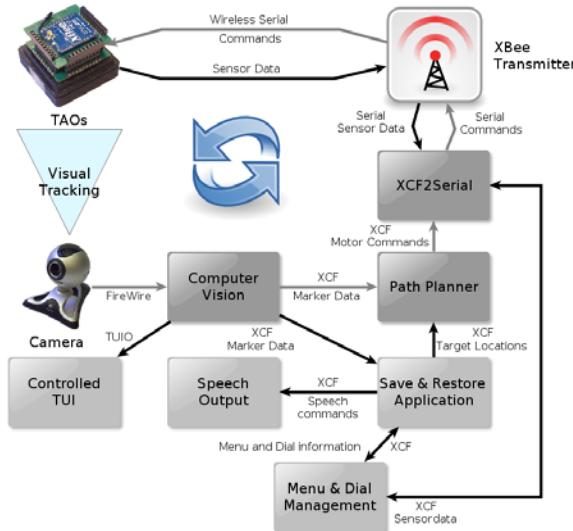


Fig. 2. The software modules and their communication relations build a closed control loop

marker information is not passed to the underlying TUI since their position is not relevant neither for the menu controlling nor the TUI. This TAO class can be easily slid over the surface by the user and even other moving TAOs. In the following we refer to these TAOs as Menu TAOs (mTAOs).

For a tangible representation of TAO arrangements we propose to apply the concept of physical containers, introduced in the mediaBlocks project (12). Menu TAOs with a dial can be associated with the arrangement of the iTAOs on the table by selecting the “save” command from the dial menu. Of course it is also possible to restore an arrangement associated with a mTAO by selecting the “restore” command. The menu is context sensitive depending on the association state of the actual mTAO. An unassigned mTAO has only two menu items: “Empty” is in the left half of the rotation range, and “save” is in the right half. If the mTAO contains a stored configuration it possesses a different menu: “Charged”, “reconstruct”, and “delete” cover each one third of the dial’s rotation range.

Fig. 4 depicts the two context-dependent menu structures. The motorized potentiometer allows to represent the association state of the mTAO. Depending on whether the mTAO has an associated arrangement or not, the menu has the items of the unassigned or assigned state and is physically set to the “empty” or “charged” item when the mTAO is put on the table-top surface.

This setup allows extending existing TUIs with the ability to manage the arrangements of all involved TUIOs, as long as the functionality of these TUIOs is static and not changeable during interaction.

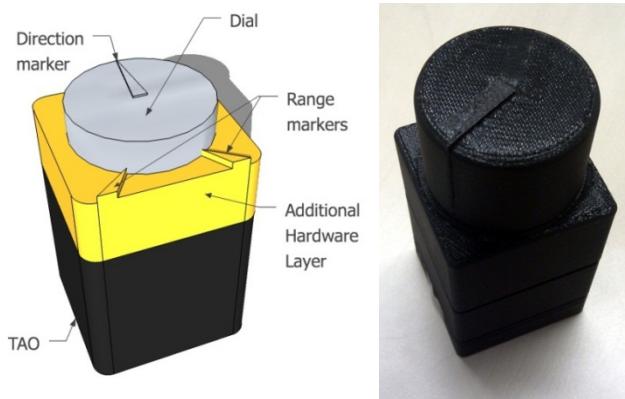


Fig. 3. Assembled second TAO type with menu dial (mTAO). The right picture shows a first working prototype of this new hardware component.

4.1 Integration of the Active TAO Menus into Passive TUIs

In order to integrate our approach into any existing TUI incapable of saving and restoring arrangements, the marker information of the TAOs needs to be made available to the control system. For this the TAO vision module offers output of the marker information using the TUO protocol described in (13) which is already used by many TUIs. In this case there is no need for an additional camera and the positions are exactly the same in the extended TUI and the TAO system.

As depicted in Fig. 2, the save-and-restore module can be invoked to save the current positions of the TAOs and also to restore them by navigating the TAOs to new target locations. To control these actions, the users simply put an mTAO on the tabletop surface. If the TAO does not carry an assigned configuration, the dial automatically rotates left to represent the “empty” state according to the menu layout depicted in Figure 4. To save a configuration, the user rotates the dial rightwards to the “save” menu item. Thereby the mTAO physically represents that it now carries a configuration and the new menu layout as shown on the right in Fig. 4 is active with respect to the new state. Verbal feedback indicates successful state changes. Correspondingly the dial rotates right to the “charged” state if a mTAO carrying a configuration is put on the table. To reconstruct the arrangement the user rotates the menu dial to the “reconstruct” menu item. The iTAOs then start to navigate to the new target locations and the TAO dial rotates back to the “charged” state. The users also have the possibility to delete the content of a menu TAO by rotating the menu dial left to the “delete” menu item. Thereby the “empty” state is directly represented. State changes and triggering commands by selecting the particular menu item also results in vocal feedback.

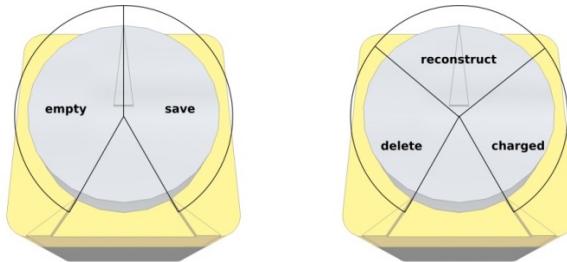


Fig. 4. Menu layouts for the “empty” and the “charged” state

5 Application Examples

As proof of concept we demonstrate our system in conjunction with an existing TUI for an interactive sonification of traffic data. Furthermore we propose a second application for family therapist which we consider for follow-up implementation.

5.1 Working with a Tangible Interface for an Interactive Sonification

As first application, we demonstrate our approach to save and restore TUIs at hand of SoundBlox³ project, a TUI for the interactive sonification of traffic data, or more precisely, a vehicle registration station at a highway where for each car, bus or lorry certain features such as time, velocity etc. are recorded. Sonification is the auditory representation of data, allowing for instance monitoring of processes without the need to attend any visual display. In SoundBlox, TUIOs represent acoustic parameters such as pitch, level or brightness of short sound events. On the desk surface, specific fixed locations are defined to represent features of the data set such as the vehicle velocity, vehicle type, distance to the vehicle before, etc. By arranging the TUIOs on the desk, the user can now interactively control how data features are represented by sound features in the real-time rendered data sonification. For instance, moving the pitch object closer to the velocity location strengthens the connection so that faster vehicles become audible as higher pitched sounds. In search of suitable sonifications, where for instance traffic jams and fluent traffic can be better discerned, users wish to record the object arrangement as a preset to be used later, e.g. to compare the sound of different configurations. To this end, our mTAO can easily be applied. Since the mTAO can be passed between users, this allows several users to work together. The system even allows multiple mTAOs to store several presets at the same time. A video demonstration of this application is provided on our website.⁴

³ <http://tangibleauditoryinterfaces.de/index.php/tai-applications/audiodome-soundblox/>

⁴ <http://www.techfak.uni-bielefeld.de/ags/ami/publications/RHR2011-SRM/>

5.2 Therapist Assistance for Measuring Family Relationships

As promising application area we propose to use our approach for the problem of measuring family relationships as described in the Family-System-Test (FAST) by Gehring et. al. (14). In this method patients use a board subdivided in 81 square fields and figures distinguishable by gender with dots as eyes and additional cubes for different heights of the figures to describe their mental view of their family relationship. We suggest to use our iTAOs as physical representation of family members. Adding 'shoulders' and a 'nose' to the squared objects even allows to use relative orientation of the objects to other objects. Additional TUImod layers allow changing the height of the iTAOs. Therapist and patient can jointly setup a configuration which fits and the mTAO could be used to save a configuration for later review, or to restore another configuration constructed from another perspective, e.g. the family system as experienced in childhood. The therapist can then use the mTAO as described above to save the arrangement and load it for later therapy sessions or analysis.

6 Discussion and Conclusion

In this paper we have presented a system that allows enhancing table-top TUIs with saving and restoring capabilities through a fully tangible menu metaphor. This approach allows multiple users working with the TUI to access menus equally. Our approach allows to have multiple mTAOs on the table to enable the comparison of multiple arrangements.

By using motorized potentiometers the state of the menu dial itself can be saved and restored as an inner degree of freedom for our actuated tangible objects. This is useful in future developments.

Our system is in an early state so that there is still room for technical improvements. Here for we currently create a new tracking system with custom markers that use infrared light. Furthermore we consider recording and segmenting interaction sequences and add further interaction possibilities known from GUIs such as undo and redo functions.

We also plan to implement the second application proposed in this paper. Furthermore we want to conduct a user study that proves the usability of these applications.

Acknowledgements. We thank the German Research Foundation (DFG) and the Center of Excellence 277 Cognitive Interaction Technology (CITEC) who funded this work within the German Excellence Initiative.

References

1. Patten, J., Recht, B., Ishii, H.: Audiopad: A tag-based interface for musical performance. In: Proceedings of the 2002 Conference on New Interfaces for Musical Expression, pp. 1–6. National University of Singapore, Singapore (2002)
2. Jordà, S., et al.: The reactable*. In: Proceedings of the International Computer Music Conference (ICMC 2005), Barcelona, Spain, pp. 579–582 (2005)

3. Han, J.Y.: Low-cost multi-touch sensing through frustrated total internal reflection. In: Proceedings of the 18th Annual ACM Symposium on User Interface Software and Technology, p. 118. ACM, New York (2005)
4. Brave, S., Ishii, H., Dahley, A.: Tangible interfaces for remote collaboration and communication. In: Proceedings of the 1998 ACM Conference on Computer Supported Cooperative Work, pp. 169–178. ACM, New York (1998)
5. Pangaro, G., Maynes-Aminzade, D., Ishii, H.: The actuated workbench: computer-controlled actuation in tabletop tangible interfaces. In: Proceedings of the 15th Annual ACM Symposium on User Interface Software and Technology, pp. 181–190. ACM, New York (2002)
6. Weiss, M., et al.: Madgets: actuating widgets on interactive tabletops. In: Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology, pp. 293–302. ACM, New York (2010)
7. Rosenfeld, D., et al.: Physical Objects as Bidirectional User Interface Elements. In: IEEE Computer Graphics and Applications, pp. 44–49. IEEE Computer Society, Los Alamitos (2004)
8. Riedenklau, E.: TAOs - Tangible Active Objects for Table-top Interaction. [Master's thesis]. Faculty of Technology, Bielefeld University, Bielefeld, Germany (2009)
9. Hermann, T., Henning, T., Ritter, H.: Gesture Desk – An Integrated Multi-modal Gestural Workplace for Sonification. In: Camurri, A., Volpe, G. (eds.) GW 2003. LNCS (LNAI), vol. 2915, pp. 369–379. Springer, Heidelberg (2004)
10. Bovermann, T., et al.: TUImod: Modular Objects for Tangible User Interfaces. In: Proceedings of the 2008 Conference on Pervasive Computing (2008)
11. Fritsch, J., Wrede, S.: An Integration Framework for Developing Interactive Robots. In: Brugali, D. (ed.), vol. 30, pp. 291–305. Springer, Berlin (2003)
12. Ullmer, B., Ishii, H., Glas, D.: mediaBlocks: physical containers, transports, and controls for online media. In: Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques, pp. 379–386. ACM, New York (1998)
13. Kaltenbrunner, M., et al.: TUOI: A protocol for table-top tangible user interfaces. In: Proc. of the The 6th International Workshop on Gesture in Human-Computer Interaction and Simulation (2005)
14. Gehring, T.M., Wyler, I.L.: Family-system-test (FAST): A three dimensional approach to investigate family relationships. In: Child Psychiatry and Human Development, vol. 16. Springer, Heidelberg (1986)
15. Patten, J., Recht, B., Ishii, H.: Audiopad: A tag-based interface for musical performance. In: Proceedings of the 2002 Conference on New Interfaces for Musical Expression, Singapore, pp. 1–6 (2002)

Needle Insertion Simulator with Haptic Feedback

Seungjae Shin^{1,2}, Wanjoo Park¹, Hyunchul Cho¹,
Sehyung Park¹, and Laehyun Kim¹

¹ Korea Institute of Science and Technology, Intelligence and Interaction Center,
39-1, Hawolgok-dong, Seongbuk-gu, Seoul, 136-791, Korea

² University of Science and Technology, HCI & Robotics department,
217, Gajung-ro, Yuseong-gu, Daejeon, 305-350, Korea
yycome@gmail.com, {wanjoo, hccho, sehyung, laehyunk}@kist.re.kr

Abstract. We introduce a novel injection simulator with haptic feedback which provides realistic physical experience to the medical user. Needle insertion requires very dexterous hands-on skills and fast and appropriate response to avoid dangerous situations for patients. In order to train the injection operation, the proposed injection simulator has been designed to generate delicate force feedback to simulate the needle penetration into various tissues such as skin, muscle, and blood vessels. We have developed and evaluated the proposed simulator with medical doctors and realized that the system offers very realistic haptic feedback with dynamic visual feedback.

Keywords: Needle insertion, Medical simulation, Haptic feedback.

1 Introduction

Medical training has become a hot issue due to the evolution of medical technology and an increase of interest in health and improving the quality of life.

Injection is one of the most basic skills used in medical care. Injection training, however, was difficult to learn due a lack of experimental objects to practice on. The objects for experimentation were animals, corpses, or patients. For experimentation using animals, animal anatomy was different from that of people and there were ethical problems. When using the corpses, there was problem which was the physiological response has not gotten enough. And in experimentation with patients, the safety of the patients was not guaranteed.

Various, advanced injection simulators have been developed recently to solve these problems. Immersion Corp developed Virtual I.V [1] which is an intravenous simulator. Virtual I.V provides haptic feedback during intravenous injection process and virtual reality for injection experimentation. But it has two limitations that it cannot provide haptic feedback during removal of the needle and it do not use real catheter. Smith [2] and Zorcolo [3] presented the Catheter Insertion Simulation using a Phantom haptic interface [4], which is also made by Immersion Corporation. Phantom's advantage is that it is easy to use and the cost of the hardware is reasonable. However, Phantom does not offer enough haptic feedback due to a different purpose and design.

The needle insertion simulator we propose provides a practical injection training environment. Unlike previous haptic devices, our needle insertion simulator is

designed to offer an optimized environment for injection training and high quality haptic feedback without any loss or distortion of haptic feedback because the motor power is not transmitted by gears or belts but directly by motor to rollers.

The overall system of the needle insertion simulator is shown in figure 1.



Fig. 1. The overall system of the needle insertion simulator

2 Hardware Configuration

This section covers the details of the hardware configuration. The simulator's hardware consists of two main parts: one is the controlling part to calculate the values of the haptic feedback and control the motor, the other is the driving part, which contains the motor and inserted hole.

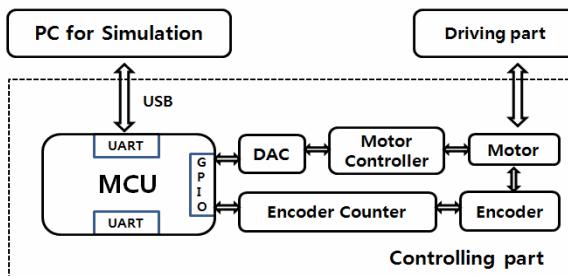


Fig. 2. Framework of the controlling part

2.1 Controlling Part

The block diagram in figure 2 illustrates the overall framework of the controlling part. The main processor of the controlling part is the DSP TMS320F280 manufactured by Texas Instruments, and the Digital to Analog Convertor (DAC) is the DAC0800 obtained from National Semiconductor. We used an RE25 DC motor which works at

24 V/0.6 A and 28.8 mN torque. The MR128 encoder attached to the motor has a 1000 pulse resolution per cycle and this 1000 pulse resolution is increased to four times by the encoder counter's quadrature counting mode, so that it has a total of 4000 resolution per cycle. Also we mounted a 4-Q-DC for the motor controller and SA1 for the tilt sensor which was obtained from DAS Corporation.

The controlling part is explained below. When a needle is inserted into the driving part, the motor rotates as long as the inserted distance. Then the encoder reads the values of the distance and the measured value is delivered to the MCU. Based on the value, the MCU, which calculates the haptic feedback, drives the motor at a 10 KHz frequency. Finally, the haptic feedback is converted to an analog signal by DAC so that the response of the analog signal is faster than the response of the digital signal.

In addition, the tilt sensor offers a slope of the driving part, and this value is transferred to the PC through the USB communication with the encoder value, and used to show the position of the virtual needle.

2.2 Driving Part

The driving part is designed with five conditions to offer an optimized environment for needle insertion simulation.

1. Eliminate loss or distortion of haptic feedback
2. Minimize primitive friction
3. Minimize slipping of the needle
4. Two degrees of freedom
5. Initial insertion slope and variation range of the slope

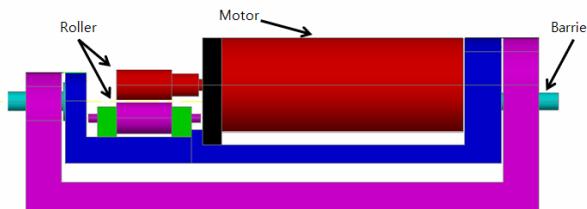


Fig. 3. Design of the driving part

The driving part satisfying the above conditions is shown in figure 3. The most important condition of the needle insertion simulator is that eliminating loss or distortion of haptic feedback. The existing haptic devices are using gears or belts to transmit the motor power. But when using gears, there is back-rash by crack between the gears, and when using belt, there is distortion of haptic feedback because the belt can be stretch. Thus, we designed new mechanism to generate delicate force feedback to simulate the needle penetration into various tissues such as skin, muscle, and blood vessels. Proposed simulator uses friction between two rollers and the needle to transmit haptic feedback. One of the rollers is connected to the motor directly. That is to say, when the needle is inserted to the hole between the rollers, the roller which is connected to the motor generates force feedback. This mechanism using friction does not have any loss or distortion of haptic feedback within maximum static frictional force.

When we insert a needle into the skin, the feeling is weak and delicate. In case of the proposed simulator, therefore, the primitive friction of rollers should be minimized. We used minimum number of bearings due to the direct connection between the motor shaft and roller for minimizing of the primitive friction. Also, the diameter of the roller was determined by an acceptable range of encoder resolution.

The rollers were not only designed to control the space between the two rollers to adjust the friction but they were also coated with urethane to prevent slipping of the needle when it is inserted. But it was hard to eliminate slipping of the needle completely because of such features as thinness and slickness. Therefore, we made a broad and furrowed needle which is shown in figure 4.

During the injection, the needle needs 2 degrees of freedom which are back, forth, up and down. In other word, the rollers should be able to lean in conformity with the needle. For this, the driving part was divided into a fixed external part and a moving internal part.

With a barrier, we could restrict the variation of the slope and adjust the initial insertion slope.

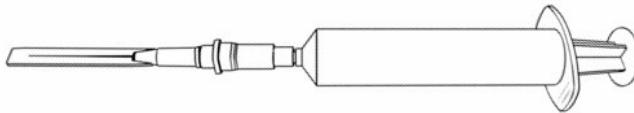


Fig. 4. The broad and furrowed needle

3 Haptic Feedback Profile

We chose intravenous injection toward antecubital basilica vein for needle insertion simulation among several injections. Antecubital basilica vein which is on the forearm is often used for intravenous injection.

In case of intravenous injection simulation, it is important to offer well matched haptic feedback when the needle passes through skin, muscle, and blood vessels. During medical treatment, we assume several conditions such as elasticity of skin, the feeling when the skin is punctured, the feeling when the needle passes through muscle, and the feeling when a blood vessel is punctured.

DiMaio [5] provided a method for quantifying the needle forces and soft tissue deformations that occur during general needle trajectories in multiple dimensions. Okamura [6] studied the force modeling for needle insertion into soft tissue. She measured the axial forces of the needle during insertion into and removal from bovine liver, and modeled force algorithm like tissue stiffness, friction, cutting force, and so on.

We defined haptic feedback as three stages for intravenous injection simulation.

1. Elasticity of skin
2. Friction in muscle layer
3. Elasticity of blood vessel

Because of the elasticity, the skin and the blood vessel have reaction against the needle insertion. But if the injection force exceeds the critical point, the needle penetrates them. Its effect is illustrated by the “Jog effect” equation 1. The effect of the penetration could be expressed by the difference between two effects as shown in figure 5.

$$f_T(T) = (T - n)^2 / a + b \quad (1)$$

After penetration, there is friction in muscle layer. We used Friction Cone Model [7] to describe this effect. The friction is calculated by equation 2. Where P_{curr_f} is the current position, P_{now} is the angular position, P_{prev} is the previous position, S_{f1} and S_{f2} are the scaling factors, P_{diff} is the difference of position, L_f is the friction level, and T_f is the friction torque.

$$\begin{aligned} L_f &= \exp(P_{now}(n)/S_{f1}) \\ P_{curr_f}(n) &= P_{prev}(n-1) + (P_{now}(n) - P_{prev}(n-1)) \times S_{f2} \\ P_{diff}(n) &= (P_{now}(n) - P_{curr_f}(n)) \times L_f \\ P_{prev}(n) &= P_{curr_f}(n) \\ \text{if } P_{diff}(n) > T_{f_max} \text{ then } T_f(n) &= T_{f_max} \\ \text{else if } P_{diff}(n) < T_{f_min} \text{ then } T_f(n) &= T_{f_min} \\ \text{else } T_f(n) &= P_{diff}(n) \end{aligned} \quad (2)$$

The elasticity of blood vessels is also described by the “Jog effect” in equation 1, but we used a smaller value of ‘a’ because the blood vessel has stronger elasticity than the skin.

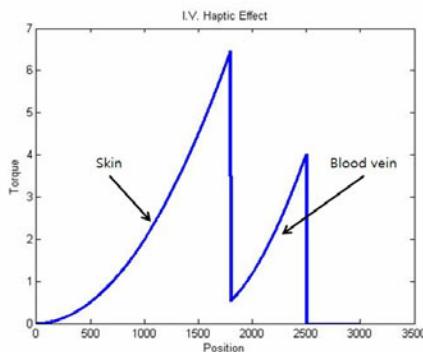


Fig. 5. Profiled haptic feedback for intravenous injection simulation

4 Results

This section covers an interview with an expert about the proposed needle insertion simulator.

We had an interview with an expert, who is an Associate Professor at the school of dentistry of Seoul National University, to evaluate the practical environment that the needle insertion simulator offers. The interview proceeded for two hours with an evaluation and a consultation after enough testing of the simulator. According to the

expert, the proposed simulator offered more advanced haptic feedback than other haptic simulators. Especially, it has advantages in illustrating the elasticity of skin, blood vessels, and feeling about the puncture. On the other hand, she advised that the distances between the skin, muscle, and blood vessels were a little bit different, and suggested a haptic needle insertion simulator combined with a model arm to improve reality.

5 Conclusions

Recently, the medical experimentation which is one of the important courses for training medical workers has become a hot issue. There were two major problems in medical experimentation, about objects and subjects. The medical simulators with haptic feedback and virtual reality have been developed recently for solving these problems.

We proposed a needle insertion simulator among various medical experimentations. Unlike previous haptic devices, the proposed simulator provides a more realistic experimentation environment because it was designed for needle insertion simulation. Also, it offers high quality haptic feedback like real feeling of the injection according to an expert. In the future, we intend to develop haptic feedback and injection experimentation contents for different ages and various parts of the human body as well as 3D virtual reality to improve the needle insertion simulator.

References

1. Laerdal Corp.,
<http://www.laerdal.com/doc/6473945/Virtual-I-V.html#>
2. Smith, S., Todd, S.: Collaborative evaluation of a haptic-based medical virtual environment. In: 4th INTUITION Int. Conf. on Virtual Reality and Virtual Envir., Athens, pp. 102–110 (October 2007)
3. Zorcolo, A., Gobbetti, E., Pili, P., Tuveri, M.: Catheter Insertion Simulation with Combined Visual and Haptic Feedback. In: 1st PHANTOM Users Research Symposium (May 1999)
4. Massie, T., Salisbury, K.: The PHANTOM Haptic Interface: A Device for Probing Virtual Objects. In: The ASME Winter Annual Meeting, Symposium on Haptic Interfaces for Virtual Envir. And Teleoperator Systems, Chicago, IL (November 1994)
5. DiMaio, S., Salcudean, S.: Needle insertion modeling and simulation. In: IEEE Int. Conf. Robotics Automat., vol. 2, pp. 2098–2105 (2002)
6. Okamura, A., Simone, C., O’Leary, M.: Force modeling for needle insertion into soft tissue. IEEE Trans. on Biomedical Engineering 51(10) (October 2004)
7. Melder, N., Harwin, W.: Extending the Friction Cone Algorithm for Arbitrary Polygon Based Haptic Objects. In: 12th Int. Symposium on Haptic Interfaces for Virtual Envir. and Teleoperator Systems, pp. 234–241 (2004)

Measurement of Driver's Distraction for an Early Prove of Concepts in Automotive Industry at the Example of the Development of a Haptic Touchpad

Roland Spies¹, Andreas Blattner², Christian Lange¹, Martin Wohlfarter¹, Klaus Bengler², and Werner Hamberger³

¹ Ergoneers GmbH, Mitterstraße 12, 85077 Manching

² Institute of Ergonomics, Technical University of Munich,

Boltzmannstraße 15, 85747 Garching

³ AUDI AG, Development HMI, 85045 Ingolstadt

{spies,lange,wohlfarter}@ergoneers.com,

{blattner,bengler}@lfe.mw.tum.de,

werner.hamberger@audi.de

Abstract. This contribution shows how it is possible to integrate the user's behavior in the development process in a very early stage of concept. Therefore innovative applied methodologies for objectifying human behavior such as eye tracking or video observation like the Dikablis/ DLab environment in the Audi driving simulator are necessary. A demonstrative example therefore is the predevelopment of a touchpad with an adjustable haptic surface as a concept idea for infotainment interaction with the Audi MMI. First an overview of the idea of capturing human behavior for evaluating concept ideas in a very early stage of the development process is given and how it is realized with the Dikablis and DLab environment. Furthermore the paper describes the concept idea of the innovative control element of the haptic touchpad as well as the accompanied upcoming demands for research and how these questions were clarified. At the end some example results are given.

Keywords: Eye Tracking, haptic feedback, touchpad, interaction, driving simulator.

1 Introduction

In future, successful products have to focus on ergonomics, usability and joy of use in order to differentiate from their competitors. Nowadays, this is one of the automobile manufacturer's biggest challenges considering the development of infotainment features. But those key factors cannot be tested quickly at the end of the development process to check if they are fulfilled. They need to be considered carefully from the very beginning of the concept idea along the whole development process. This assumes an integration of theoretical ergonomic product design as well as a consideration of human skills, behavior and user's opinion. Especially for the last point lots of methodologies such as e.g. focus groups or standardized questionnaires do exist.

For a complete user centered product design objective behavioral data are also necessary. Reasons for that are documents like the AAM [1] for example, which contain specific criteria of driving behavior and gaze distraction for safety reasons which have to be fulfilled before entering the market.

Furthermore at the very beginning of the development process prototypes are on a lower level than the latest products in market concerning design and functionality, which have a massive impact on subjective justification. Bringing design and functionality issues to perfection is very costly. Behavioral strategies like e.g. driving performance or gaze distribution should be unaffected by design issues.

The following chapter gives an overview about the basic ISO standards for capturing the required gaze metrics as well as how this is implemented in the Dikablis DLab environment for an efficient and effective evaluating process.

Afterwards the idea of developing a touchpad with an adjustable haptic surface for in-vehicle infotainment control for the Audi MMI is presented. This example shows the evaluation of an ergonomic concept for the decision making of a further complex technical development.

2 Behavioral Research According to ISO - Standards

The synchronous recording and analyzing of data is always a challenge in experiments. This is especially the case when eye-tracking and data of additional data streams such as e.g. external videos shall be recorded together with driving relevant data. One requirement for a successful test is the planning part at the beginning. To face this challenge, the D-Lab environment [2] is used in the Audi driving simulator. D-Lab enables synchronous recording and common analysis of several data streams such as eye-tracking data, several video streams, audio, driving dynamics data, physiological data and workload measurement. Furthermore D-Lab contains a build-in planning, measure and analyzing process according to the process described in ISO/TS 15007-2:2001 [3] which guides the experimenter through the experimental process. Below these three steps are described shortly.

2.1 Planning an Experiment

ISO/TS 15007-2:2001 [3] describes how behavioral experiments shall be planned. The idea is to think about the tasks, which will be given to the subject during the experiment and to cut major tasks further down. The highest level is the so called condition (e.g. a new navigation system). This condition can contain tasks (e.g. enter navigation destination) and those tasks can contain subtasks (e.g. enter town, street and house number). Figure 1 on the left shows the suggestion of ISO/TS 15007-2:2001 and Figure 1 on the right shows how one can set up this experimental plan in D-Lab according to the suggestion of ISO/TS 15007-2:2001.

2.2 Synchronous Measurement with D-Lab

The core when performing an experiment is D-Lab. D-Lab allows to control the experiment and to record all data synchronously. In order to control the experiment

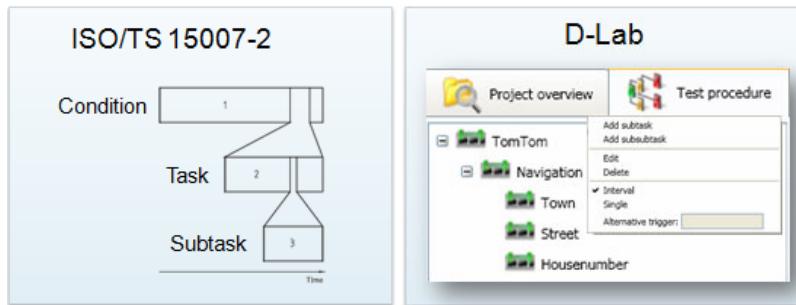


Fig. 1. Left: Experimental plan according to ISO/TS 15007-2:2001; Right: Planning module in D-Lab

D-Lab visualizes the experimental plan as clickable buttons. These buttons can be clicked to mark beginning and end of task intervals. Clicking of those buttons can be done with the mouse as well as by network events.

When the recording has been started D-Lab records synchronously all connected measurement devices like in this case glance behavior via Dikablis, driving performance data from the Audi driving simulator, video data from 4 cameras and subject's comments via a microphone. The interconnection is shown in Figure 2.

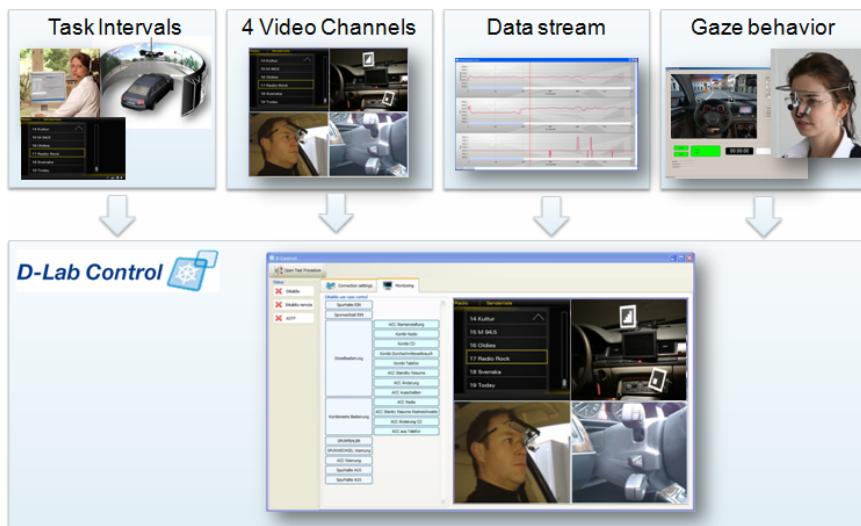


Fig. 2. Experimental control and synchronous data measurement with D-Lab

The great advantage of the D-Lab research environment is that all data is recorded synchronously and already contains the intervals of all performed tasks. This simplifies and shortens the data analysis after the experiment extremely. Figure 3 shows the common replay of gaze data, driving dynamics data and the 4 recorded videos in D-Lab.

2.3 Common Data Analysis in D-Lab

All synchronously recorded data of the whole project can be imported and analyzed with D-Lab. After all recorded data is imported one can replay the whole dataset synchronously and analyze it step by step (see Figure 3). The analysis process is described below using the example of analyzing eye-tracking data. The data analysis of all other measured data follows exactly this analysis process.

For the analysis of the eye-tracking data, D-Lab offers the possibility for the free definition of areas of interest (AOIs), for which the glance durations are calculated automatically and conforming to the eye-tracking standard ISO/TS 15007-2. The glance durations are shown as timeline bars synchronously to the progress bar of the gaze movie respectively to the progress bar of the player for the four recorded video channels (see Figure 3). The vertical line under the gaze movie player shows exactly the progress of the gaze movie or the four external videos in the timeline bars of the glance durations to the AOIs.

With the help of the integrated statistics functionality can be defined, which glance metrics shall be calculated for which AOIs during which task. An example for such a calculation would be: Calculate the glance metrics "total glance time", "total glance time as a percentage" and "number of glances" to the AOI display of the navigation system while entering the navigation destination. D-Lab calculates these metrics automatically and visualizes the result in a table for all subjects of an experiment. This automated calculation can be done for all defined AOIs, all experimental conditions, tasks and subtasks as well as for all glance metrics. The calculated result which is visualized in a table can be exported to a .csv file which can be imported to SPSS or MS Excel.

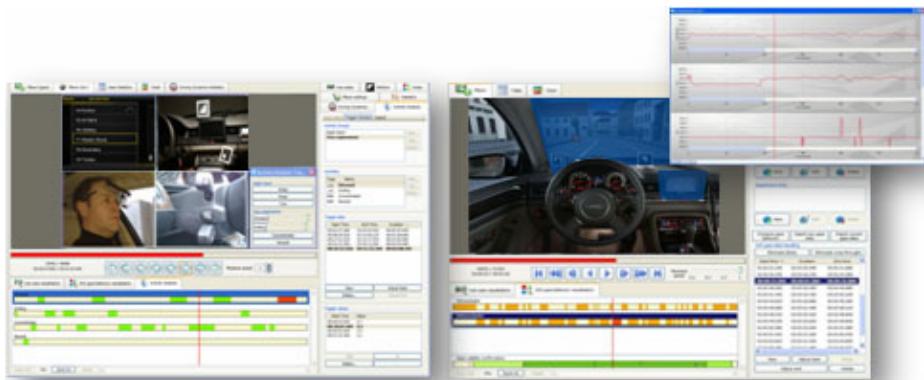


Fig. 3. Synchronous replay and common analysis of all data in D-Lab

3 Haptic Touchpad for Infotainment Interaction

The following chapter contains the concept idea of the Haptic Touchpad for infotainment interaction. The technical realization for an automotive application of such a device is very costly and complicated. The idea was to search for a technology

which enables the functionality independently from a later technical realization in a car application. The goal for such a prototype was to evaluate the theoretical hypothesis and prove the estimated benefit of such a control device.

3.1 Concept Idea and Technical Realization of a Haptic Touchpad

A challenge for car manufacturers today is the increasing amount of comfort functions in modern vehicles e.g. navigation, media and communication systems and soon internet services. To keep all these systems controllable while driving, car producers have integrated these functions in menu-based central infotainment systems which are mostly controlled by one multifunctional input device. Currently, many different solutions of such control devices are available on the market. These solutions can be divided into two different groups: First, an integrating approach represented by touchscreens and second an approach separating display and control element, e.g. turning knobs or joysticks in the center console. The latest concept with an additional touchpad for character input is delivered by Audi [4]. According to Hamberger [5] an in-car-touchpad offers several potentials. It is familiar to the users because of the accustomed usage with computer touchpads and enables handwriting recognition. Furthermore a touchpad is a multipurpose control element, which perpetuates the possibility of splitting up display and controls. Thus the display remains in the ideal field of vision and the touchpad can be positioned in the ideal reaching distance. Robustness, optics and the ease of use are additional positive arguments. The results of an experiment in a driving simulator occupy, that a touchpad reduces track deviation compared to a rotary push button and a touchscreen. Within the task of a text entry a touchpad decreases gaze diversion times in comparison to a rotary push button. In addition to the mentioned results customers prefer the usage of a touchpad compared to a touchscreen [5].

In order to enable the control of the whole infotainment menu with one single touchpad, the idea is to give an additional haptic structure onto the surface for orientation. A prototype of a touchpad with such an adjustable haptic structured surface has been built up for evaluation purposes [6] [7]. In order to realize the haptic surface the technology of hyper braille is used [8]. The touchpad surface can be adjusted to the displayed content on the middle screen, so that the user can feel and press every elevated element on the touchpad which is shown on the display (see Figure 4).



Fig. 4. Left: Elevated elements on the touchpad; Right: Displayed content on the middle screen with highlighted graphical widget

An additional visual feedback by highlighting the current touched graphical widget on the screen shows the user the finger position on the touchpad (see Figure 4). This guarantees an absolute compatible interface design between display and control device because of direct manipulation.

Moreover new interaction concepts, like it is suggested by Broy [9] for example, are possible by using a two dimensional touch device. A touchpad enables a separation of display and control, what means that the position of the control element is independent from the display position. Hence infotainment information can also be provided in the Head-Up Display [10] or further innovative interaction concepts as it is proposed by Spies et al. [11] are thinkable.

3.2 Need for Research

In case of menu operation while driving it is about a dual task situation which can cause interferences between the two parallel tasks. This means that menu control leads to driver distraction from the driving task. The major goal of an ergonomic interface design is to avoid distraction from driving because of controlling a secondary task like e.g. a navigation system. The ultimate benefit of the haptic structured touchpad surface is to give the user additional orientation via the haptic channel by elevating different kinds of shapes on the touchpad surface. Thus the haptic structure is not only for finding elements but also for getting the content of the element. A reduced number of glances away from the driving scene to the screen are expected.

In terms of an automotive capable realization the highest costs are produced by the realization of a technology for the haptic structured surface. To clarify if this is really worth it and if there is a real benefit given by the haptic structure, two questions should be answered:

1. Is there a real benefit given by the haptic structure compared to a usual flat touchpad?
2. Is a cursor for giving an optical feedback of the current finger position comparable to a laptop application enough?

3.3 Simulator Tests for Capturing Human Driving and Gaze Behavior

How to clarify the mentioned questions in a very early stage of concept development is shown with the results of a usability test in a static driving simulator at Audi [12]. The driving task consisted of following a vehicle in a constant distance. The test persons had to fulfill different menu tasks with the navigation system while driving. The navigation system had to be controlled via four different touchpad solutions:

1. Haptic Touchpad with cursor Feedback (HTP + Cursor)
2. Haptic Touchpad without Feedback (HTP without Cursor)
3. Flat Touchpad with Cursor (TP + Cursor)
4. Flat Touchpad without Cursor (TP without Cursor)

The gaze behavior was measured with Dikablis and was synchronously stored with driving performance data via DLab Control. Figure 5 shows the results for the gaze behavior as well as the driving performance.

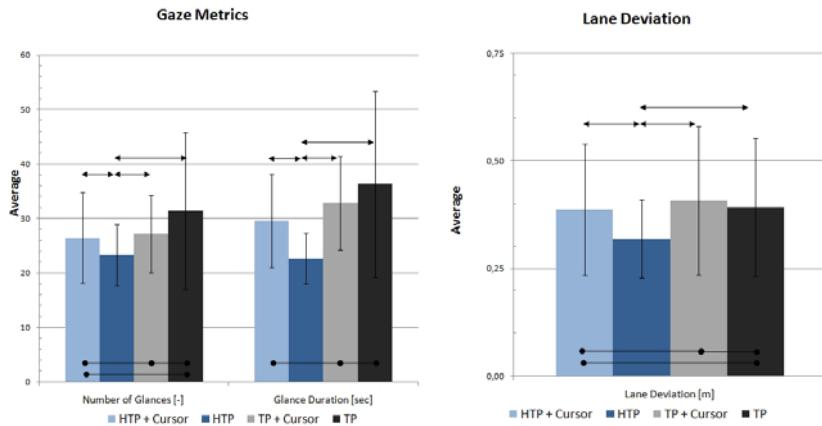


Fig. 5. Left: Number of gazes and gaze duration to the display while controlling a secondary task with different touchpad solutions; right: Lane deviation while controlling a secondary task with different touchpad solutions

The results show that the haptic structured surface improves the control performance with the touchpad significantly. The number of gazes as well as the gaze duration can be decreased by providing additional orientation feedback via the haptic channel. This also leads to an improvement of the lane deviation what comes along with safer driving performance.

The theoretical assumption would have been that an additional optical cursor feedback leads to a further improvement of driving safety and gaze distraction. The results show that this assumption cannot be verified via the objective behavioral data. Moreover the optical cursor feedback even works like an eye catcher that causes additional gaze distraction from driving.

4 Conclusion

In conclusion the example of evaluating the concept idea of the haptic touchpad in a realistic environment in a very early stage of development shows, how on the one hand theoretical ideas can be proven if they work under realistic conditions like it has been done with the haptic surface. In this case the haptic structure leads to a massive benefit what legitimates a further effort in development of the technical realization. If the result would have been, that there is no difference between haptic and flat touchpad surface the technical development could have been stopped at this point what would have avoided the wastage of a high amount of development budget. The example with the optical feedback shows how the measurement of objective behavioral data can avoid wrong concepts given by theoretical assumptions.

References

1. AAM, Driver Focus-Telematics Working Group: Statements of Principles, Criteria and Verification Procedures on Driver Interactions with Advanced In-Vehicle Information and Communication Systems. Alliance of Automobile Manufacturers (2006)

2. Lange, C., Spies, R., Wohlfarter, M., Bubb, H., Bengler, K.: Planning, Performing and Analyzing eye-tracking and behavioral studies according to EN ISO 15007-1 and ISO/TS 15007-2 with Dikablis & D-Lab. In: Proceedings of the 3rd International Conference on Applied Human Factors and Ergonomics, Miami, Juli 17-20 (2010)
3. ISO/TS 15007-2:2001 – Road vehicles – Measurement of driver visual behaviour with respect to transport information and control systems – Part 2: Equipment and procedures
4. Hamberger, W., Gößmann, E.: Bedienkonzept Audi: Die nächste Generation. In: VDI Wissensforum GmbH (Hrsg.) Elektronik im Kraftfahrzeug. VDI-Berichte, vol. 2075, pp. 677–686. VDI-Verlag, Düsseldorf (2009)
5. Hamberger, W.: MMI Touch – new technologies for new control concepts. In: IQPC – Automotive Cockpit HMI 2010, Steigenberger Graf Zeppelin, Stuttgart (2010)
6. Spies, R., Peters, A., Toussaint, C., Bubb, H.: Touchpad mit adaptiv haptisch veränderlicher Oberfläche zur Fahrzeuginfotainmentbedienung. In: Brau, H., Diefenbach, S., Hassenzahl, M., Kohler, K., Koller, F., Peissner, M., Petrovic, K., Thielsch, M., Ullrich, D., Zimmermann, D. (Hrsg.) Usability Professionals. Fraunhofer Verlag, Stuttgart (2009)
7. Spies, R., Hamberger, W., Blattner, A., Bubb, H., Bengler, K.: Adaptive Haptic Touchpad for Infotainment Interaction in Cars – How Many Information is the Driver Able to Feel? In: AHFE International – Applied Human Factors and Ergonomics Conference 2010. Wiley-Blackwell, Oxford (2010)
8. Hyperbraille (2010), <http://www.hyperbraille.de/>
9. Broy, V.: Benutzerzentrierte, graphische Interaktionsmetaphern für Fahrerinformationssysteme. Technischen Universität München, Dissertation (2007)
10. Milicic, N., Platten, F., Schwalm, M., Bengler, K.: Head-Up Display und das Situationsbewusstsein. In: VDI Wissensforum GmbH (Hrsg.) Der Fahrer im 21. Jahrhundert Fahrer, Fahrerunterstützung und Bedienbarkeit. VDI-Berichte, vol. 2085, pp. 205–219. VDI-Verlag, Düsseldorf (2009)
11. Spies, R., Ablaßmeier, M., Bubb, H., Hamberger, W.: Augmented interaction and visualization in the automotive domain. In: Jacko, J.A. (ed.) HCI International 2009. LNCS, vol. 5612, pp. 211–220. Springer, Heidelberg (2009); Dendrinos, D.S.: Traffic-flow dynamics: a search for chaos. Chaos, Solitons and Fractals 4(4), 605–617 (1994)
12. Spies, R., Horna, G., Bubb, H., Hamberger, W., Bengler, K.: Haptisches Touchpad - Zentrales Infotainmentbedienteil mit adaptiv haptisch veränderlicher Oberfläche. In: GfA (Hrsg.) Neue Arbeits- und Lebenswelten gestalten. Bericht zum 56. Kongress der Gesellschaft für Arbeitswissenschaft, pp. 123–126. GfA-Press, Dortmund (2010)

A Tabletop-Based Real-World-Oriented Interface

Hiroshi Takeda¹, Hidetoshi Miyao², Minoru Maruyama², and David Asano²

¹ Graduate School of Science of Technology, Shinshu University

4-17-1 Wakasato, Nagano, 380-8553, Japan

² Faculty of Engineering, Shinshu University

4-17-1 Wakasato, Nagano, 380-8553, Japan

t09a532@shinshu-u.ac.jp,

{miyao,maruyama,david}@cs.shinshu-u.ac.jp

Abstract. In this paper, we propose a Tangible User Interface which enables users to use applications on a PC desktop in the same way as a paper and pen on a desk in the real world. Also, the proposed system is cheaper to implement and can be easily setup anywhere. By using the proposed system, we found that it was easier to use than normal application user interfaces.

Keywords: Tangible user interface, DigitalDesk.

1 Introduction

Real-world-oriented interfaces allow more intuitive operation of computers compared to keyboards and mice because the objects are touched directly. We aim to implement a real-world-oriented interface that allows users to intuitively operate computer applications on a tabletop.

Similar research includes the “DigitalDesk” by Wellner [1], “The Everywhere Displays Projector” developed by IBM [2], “PaperWindows” by Human Media Lab [3] and “UbiTable” developed by MERL [4].

In DigitalDesk a way to operate a computer on a desktop without a mouse or keyboard is proposed. The user operates the computer by touching images of computer applications projected onto his table by a projector attached to the ceiling over the table. The shape of the user's hand is determined by processing the image taken by a camera placed next to the projector. Our system is similar in that it also uses a projector to project images of the applications onto a table so that a user can intuitively operate the applications. However, while the images on the table are touched in the DigitalDesk, our system uses real paper and a pen to operate the applications. In this way, we feel that movement of the applications can be done more intuitively. Moreover, our system is cheaper and easier to implement.

The Everywhere Displays Projector is a system that allows multiple users to efficiently work by projecting GUIs to various places in a room. Distortion in the projected image results, but this can be compensated for by using a hardware accelerated projection transform. The purpose of this system is completely different than ours since we are only interested in operating applications on a table, but we also use a similar approach to speed up rendering of the projected images.

UbiTable is a system to share applications among multiple users on a table. In particular, applications in mobile appliances, e.g., notebook PCs and PDAs, can be shared on the table. Our system is similar except for the ability to handle multiple users. However, UbiTable requires the use of a special device called “DiamondTouch”[5]. Our system does not require any special equipment and can therefore be implemented with ease to acquire components.

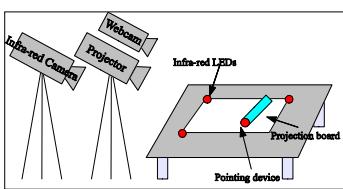
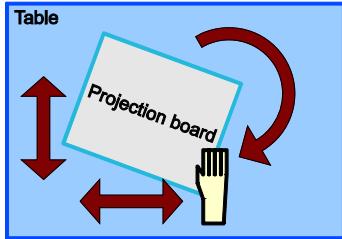
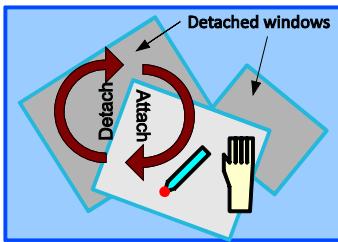
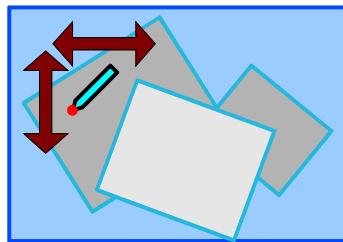
PaperWindows is a system that allows multiple users to share applications displayed on a projection board that can be moved freely around a room. This system is similar to ours in that a pen can be used to draw and click on the projection board to perform various operations. However, the projection board is flexible and can be moved around a room, so equipment to track the projection board is necessary. This makes the system more expensive and difficult to implement than ours.

2 System Overview and Proposed Features

As shown in Fig.1, a projector, infra-red camera and web camera are placed next to the desk. A projection board with infra-red LEDs is placed on the desk. A pointing device is also required. The projector is used to project the application windows on the desk, while the user operates the applications on the projection board using the pointing device.

Our system has the following features:

- As shown in Fig.2, the user can hold the projection board and move it freely. When the board is moved, the application window projected from the projector follows the board. This gives the user the impression that the window is a piece of paper and can be operated as such.
- To operate the application, we use a pointing device. When the button on the side of the pointing device is pressed, infra-red light is emitted from the LED on the tip. When the button is pressed while the pointing device is on the tabletop, it is interpreted as a mouse click, and when the button is held down and the pointing device is moved, it is interpreted as a mouse drag operation. Using the pointing device in this way on the projection board allows the user to operate the applications.
- In order to handle several applications, we implemented a function called the “detach” operation that allows the user to fix a window to the tabletop (Fig.3). To detach a window, the user clicks the “Detach” button on the menu displayed on the side of the projection board. When a window is detached, the image that was displayed on the projection board is fixed in place on the tabletop. To move the window from the tabletop back to the projection board, we provide the “attach” operation (Fig.3). To attach a window, the user moves the projection board to the window fixed on the tabletop and then clicks the “Attach” button.
- The projection board image is always displayed on top of the other windows.
- Sometimes, users may want to move a window fixed to the tabletop to another place. In order to do this, the user just has to drag the window using the pointing device (Fig.4).

**Fig. 1.** System overview**Fig. 2.** Movement operation of projection board**Fig. 3.** Attach and detach operation**Fig. 4.** Operation of detached windows

3 System Implementation and Experiments

3.1 Implementation

We implemented the above mentioned features in the following way. The necessary equipment is shown in Fig.1. The infra-red camera is an inexpensive Wiimote device. The Wiimote is limited in that it can only receive a maximum of four infra-red spots, so we use three spots to detect the projection board and one spot to detect the pointing device. By detecting these infra-red spots with the infra-red camera, we can construct the image to be projected and extract the operations to be done on the application windows.

System Calibration. We aim to create a flexible system that can be used by simply putting the equipment next to any type of desk. In order to achieve this goal, the coordinate systems of the infra-red camera, projector and tabletop must be calibrated. This calibration is done using the following procedure:

1. The web camera and tabletop are calibrated by placing a checker board on the tabletop and then taking an image with the web camera.
2. The projector and tabletop are calibrated by projecting a gray code pattern [6] on the tabletop and then taking an image with the web camera.
3. The projector and infra-red camera are calibrated by projecting several points onto the tabletop and having the user place infra-red LEDs on these points.

Estimation for the position of projection board. During the operation, real-time estimate of position and orientation of the projection board on the desk is necessary.

The estimation is carried out by detecting three infra-red spots at vertices of the projection board. Correspondence is taken between the vertices of the rectangular model and the detected spots in the image, using the information on lengths of edges of the rectangle.

Even if one spot is not observed due to occlusion, motion estimation is achieved. For the visible spots, correspondence is inferred based on simple nearest neighbor search from the positions of vertices in the previous frame. Let \mathbf{x}_i and \mathbf{m}_i ($i=1,2$) be the observed spots and the corresponding model vertices, where

$$\mathbf{x}_i = R(\theta)\mathbf{m}_i + \mathbf{t}, R(\theta) = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} = \begin{bmatrix} c & -s \\ s & c \end{bmatrix}$$

Then, the following equation is derived.

$$\Delta \mathbf{x} = \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} = \mathbf{x}_2 - \mathbf{x}_1 = R(\theta)(\mathbf{m}_2 - \mathbf{m}_1) = \begin{bmatrix} c & -s \\ s & c \end{bmatrix} \begin{bmatrix} \Delta m_x \\ \Delta m_y \end{bmatrix}$$

We can estimate θ by solving

$$\begin{bmatrix} \Delta m_x & -\Delta m_y \\ \Delta m_y & \Delta m_x \end{bmatrix} \begin{bmatrix} c \\ s \end{bmatrix} = \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix}, \theta = \tan^{-1}\left(\frac{s}{c}\right)$$

3.2 Experiments

We did experiments with our system. Fig.5 shows an example of an application window displayed on a tabletop. We confirmed that the features described above functioned properly by trying various operations. We also showed that windows could be moved in the same way as moving a piece of paper, which results in an interface that is more intuitive to use than normal computer interfaces.

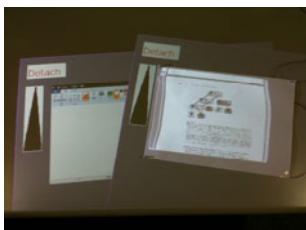


Fig. 5. Example of application windows displayed on a tabletop



Fig. 6. Environment overview

Experiments were done as shown in Fig. 6. The computer used was a Windows PC, X200(Lenovo), the camera was an easily available web camera, Live! Cam (Creative), the infrared camera was a Nintendo Wiimote and the projector was a low cost BenQ MP620C.

Moving the projection board on the tabletop. In Fig. 7 we show how to move an application attached to the projection board (Fig. 2). The projection board, on which has an application displayed, is positioned on the tabletop. The user can move or



Fig. 7. Translation and rotation of the projection board

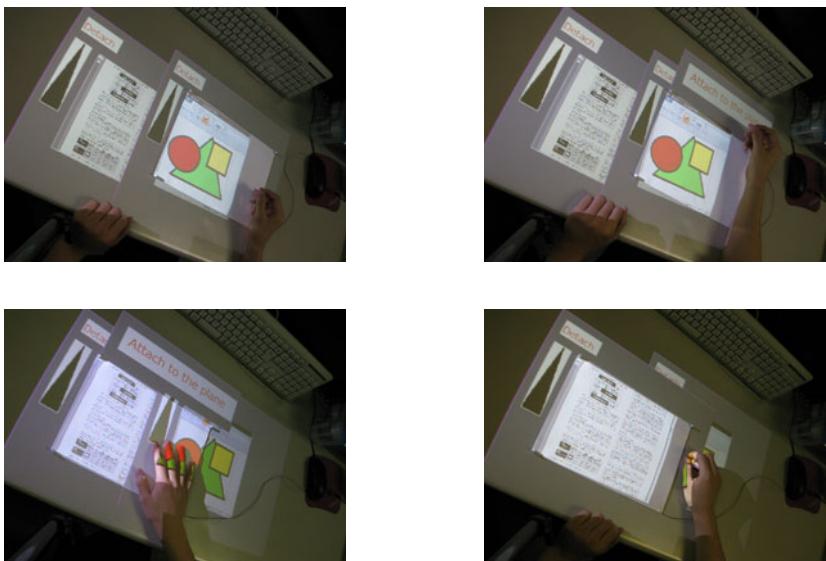


Fig. 8. Attach and detach

rotate the application by touching the projection board. If the movement is too quick, a slight delay in displaying the image results. Also, the brightness of the image depends on the distance from the projector.

Attaching and Detaching Applications. In Fig. 8 we show how to attach and detach an application and also how to change from one application to another (Fig. 3). First, the user clicks the detach button positioned at the top left of the application with the pointer. This results in no applications being attached to the projection board. In this state, an attach button is created at the top of the projection board. Second, the user moves the application he wants to attach to the projection board. Third, the user clicks the attach button to attach the desired application to the projection board. The attached application is displayed on top of any other applications and becomes the focus of further operations. Experiments with two applications showed no problems switching between them.

Moving a detached application. In Fig. 9 we show how to move an application that is not attached and is displayed below other applications (Fig. 4). The position of an application that is not attached cannot be changed by moving the projection board. To accomplish this, we use the pointer. By dragging the pointer on the tabletop, the closest application moves toward the pointer. Rotation of the application is not possible in this case. In experiments with two applications, we were able to achieve the desired results.

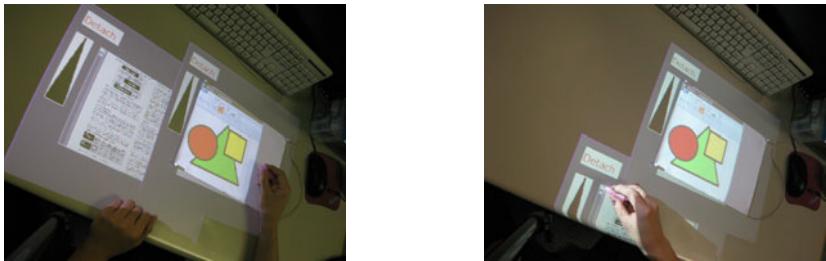


Fig. 9. Moving a detached application

4 Future Work

In future work, instead of a projection board, we plan to use a 3D type of projection object. In Fig. 10, an image being displayed on a cube is shown. In this experiment, we determined the orientation of the cube in a different way in order to render the textures. We plan to use various shaped objects in combination with the projection board.



Fig. 10. Projection onto a 3D object

5 Conclusions

In our research, we proposed and implemented the basic features for a tabletop interface using a projection board and pointing device and confirmed its operation. Our system has the advantages of being inexpensive (the necessary equipment can be purchased for about \$500) and easy to setup.

Problems that need to be examined are as follows:

- Images far from the projector are darker and have lower resolution. Therefore, text can be hard to read.
- If the projection board is moved too quickly, then the projected image may not be able to keep up.

In future work, we plan to work on methods to solve the above problems and to examine interfaces other than the projection board, such as a cube. Also, we are planning research on using the movement of a user's fingers as an input device.

References

1. Wellner, P.: Interacting with paper on the Digital Desk. *Communications of the ACM* 36(7), 87–96 (1993)
2. Pinhanez, C.: The Everywhere Displays Projector: A Device to Create Ubiquitous Graphical Interfaces. In: Abowd, G.D., Brumitt, B., Shafer, S. (eds.) *UbiComp 2001*. LNCS, vol. 2201, pp. 315–331. Springer, Heidelberg (2001)
3. Holman, D., Vertegaal, R., Altosaar, M.: Paper Windows: Interaction Techniques for Digital Paper. In: *CHI 2005*, pp. 591–599 (2005)
4. Shen, C., Everitt, K., Ryall, K.: UbiTable: Impromptu face-to-face collaboration on horizontal interactive surfaces. In: Dey, A.K., Schmidt, A., McCarthy, J.F. (eds.) *UbiComp 2003*. LNCS, vol. 2864, pp. 281–288. Springer, Heidelberg (2003)
5. Dietz, P., Leigh, D.: Diamond Touch: A Multi-User Touch Technology. In: *UIST 2001*, pp. 219–226 (2001)
6. Lee, J., Dietz, P., Aminzade, D., Hudson, S.: Automatic Projector Calibration using Embedded Light Sensors. In: *UIST 2004*, pp. 123–126 (2004)

What You Feel Is What I Do: A Study of Dynamic Haptic Interaction in Distributed Collaborative Virtual Environment

Sehat Ullah¹, Xianging Liu³, Samir Otmane¹,
Paul Richard², and Malik Mallem¹

¹ IBISC Laboratory, University of Evry,

40 rue de pelvoux 91020, France

sehat.ullah@ibisc.univ-evry.fr

<http://www.ibisc.univ-evry.fr>

² LISA Laboratory, University of Angers,

62 av Notre Dame du lac 49045, France

³ Tokyo Institute of Technology, Japan

Abstract. In this paper we present the concept of “What You Feel Is What I Do (WYFIWID)”. The concept is fundamentally based on a haptic guide that allows an expert to control the hand of a remote trainee. When haptic guide is active then all movements of the expert’s hand (via input device) in the 3D space are haptically reproduced by the trainee’s hand via a force feedback device. We use haptic guide to control the trainee’s hand for writing alphabets and drawing geometrical forms. Twenty subjects participated in the experiments to evaluate.

Keywords: Haptic guide, CVE, Virtual reality, Human performance.

1 Introduction

The Collaborative Virtual Environments(CVEs) have attained a growing interest from the research community since the last few years. A CVE is a computer generated world that enables people in local/remote locations to interact with synthetic objects and representations of other participants within it. The applications domain for such environments can be in military training, telepresence, collaborative design and engineering, entertainment and education. Interaction in CVEs can be categorized in the following way, also discussed in [36].

The environments that allow multiple users, but only one user is active at a time and is capable to interact and/or manipulate the objects. The others remain passive and wait their turn. The environments where users perceive the co-presence through avatars but each user independently interacts with the objects. Any change to the attribute of an object or scene is visible to all collaborators. The environments where two or more users can manipulate the same object. This type of manipulation can be asynchronous or synchronous. The synchronous/concurrent manipulation is also termed as cooperative manipulation.

CVEs can be used for learning purposes. For example, if one or more users watch, how a particular task is performed by the expert in the Virtual Environment (VE). This is called learning by observation but it may not be so much effective without haptic feedback because it plays a vital role in learning or transferring motor skills [19][18][16][17]. Most of these systems are single user and use the mechanism of record and play of force and position values. Similarly, force feedback has also been used in both single user and collaborative VEs in order to impart the users more realism and increase their performance but dynamique haptic guidance in 3D distributed CVEs has rarely been investigated.

This section is followed by the related work. Section 3 presents the architecture of the VE. Section 4 presents the model for the haptic guide and its evaluation. Conclusion is presented in section 5.

2 Related Work

Basdogan et al. have investigated the role of force feedback in cooperative task. They connected two monitors and haptic devices to a single machine [4]. Similarly, sallnas et al. have reported the effect of force feedback over presence, awareness and task performance in a CVE. They also connected two monitors and haptic devices to a single host [11]. But all these systems use force feedback only for realism and not for guidance in collaboration. Other important works that support the cooperative manipulation of objects in a VE include [4][5] but all these systems require heavy data exchange between two nodes to keep them consistent. Haptic guides have successfully been used for 3D object selection in large scale VEs [13]. McSig is a multimodal teaching and learning environment for visually-impaired students to learn character shapes, handwriting and signatures collaboratively with their teachers. It combines haptic and audio output to realize the teacher's pen input on a specialized sheet for visually disabled people [8]. Here both teacher and student connect to the same computer. CHASE (Collaborative Haptics And Structured Editing), is a synchronous structured drawing tool. It provides telepointers and allows users to simultaneously work on a large canvas while each maintaining a separate view of it [15]. In [7] a virtual environment that allows two users to collaboratively sculpt from remote locations, has been presented. Here haptic feedback is used to render the tool's pressure on the clay and to avoid the simultaneous editing of a vertex. Similarly Chan et al. [2] have reported the use of vibro-tactile cues to facilitate turn-taking in an environment that supports collaboration but only one user remains in control and has the rights to manipulate objects at a particular instant. Virtual Fixtures (VFs) formalism has been presented in [12] and mechanics based characterization and design guidelines have been reported in [9].

Similarly, Chellali et al. have also investigated the learning of an insertion task through haptic feedback [20]. In this study, they allowed two users to see the virtual world using two desktop displays with the same computer while sending the position values of the master Virtuose to the slave through network. The system that we propose is distributed and guidance is not limited to a single task.

3 Dynamique Haptic Interaction

3.1 System Architecture

Here we present our system that allows an expert to haptically guide the hand of non-expert in the 3D space. The two users (expert and non-expert) use two separate machines connected through local area network. Our application that allows collaboration is installed on each of these machines [14]. Whenever the applications are launched on both systems and network connection is successfully established between them, each user will see the VE on their screens. The VE has room a like structure to give the 3D perception. This environment contains two spheres each representing a user. Each user controls the movements of his/her corresponding sphere through a haptic device.

3.2 Haptic Guide

The name “What You Feel Is What I Do” is given to the concept because of the manner the force acts on the trainee. When all the conditions for haptic guide’s activation are satisfied then all movements by the expert’s input device in the 3D space are haptically reproduced by the trainee’s hand via a force feedback device (Phantom and SPIDAR in our case). More specifically any 3D motion done by the expert is reproduced and haptically felt by the trainee or the person using the second machine. The spatial information (position) of the two input devices are mutually exchanged and visualised on both machines but the force is only locally calculated on the trainee’s machine.

Referring to the figure 1, U_e and U_t represent the spheres of the expert and trainee respectively. Similarly r_e and r_t are the radii of the spheres representing the expert and trainee respectively. The distance between the two is denoted by d .

$$P_{ue} = (X_{ue}, Y_{ue}, Z_{ue}) \quad (1)$$

$$P_{ut} = (X_{ut}, Y_{ut}, Z_{ut}) \quad (2)$$

Here P_{ue} and P_{ut} represent the positions of the expert’s and trainee’s sphere respectively. In order to activate the guide, the expert envoks an event for example pressing a button of the phantom stylus. This event will change the color of the expert’s sphere on both machines. The second condition that must be true for guide activation is:

$$d \leq r_e * r_t \quad (3)$$

The guiding force which attracts the trainee’s sphere towards the centre of the expert’s sphere is calculated in the following way,

$$F = K * ((X_{ue} - X_{ut})_x + (Y_{ue} - Y_{ut})_y + (Z_{ue} - Z_{ut})_z) \quad (4)$$

Where K is a constant and was experimentally set in order to have a smooth but not very rigid force. This guide can be used for children to learn alphabets and digits of various languages (i.e english, arabic etc.), geometrical forms and

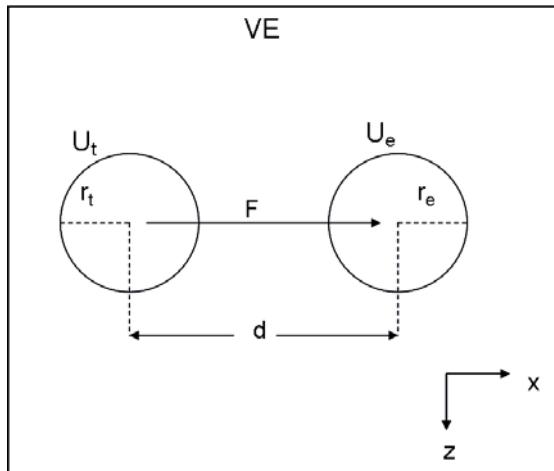


Fig. 1. Illustration of haptic guide model (*top view of VE*)

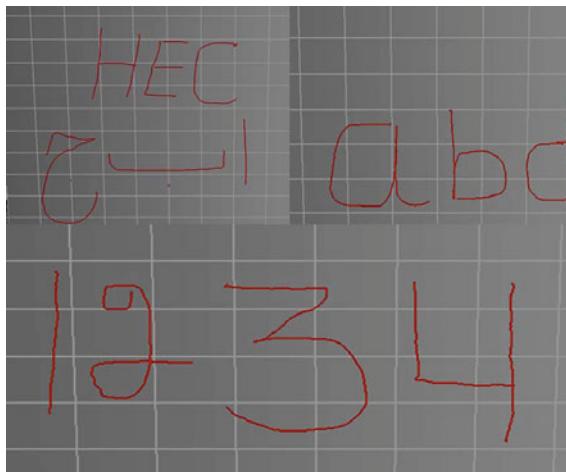


Fig. 2. Writing alphabets and digits with the help of the haptic guide

drawings (see figure 2) in the supervision of an expert (teacher). The guide does not require to record the characters, digits or geometrical forms in advance.

The concept was implemented using two different configurations. In the first implementation, we used two desktop computers connected through Local Area Network (LAN). Each computer was equipped with a Phantom omni. In the second implementation two large scale platforms were used. Both platforms are equipped with SPIDAR (Space Interface Device for Artificial Reality) (see figure 3) [21] and connected through LAN. Here, SPIDAR-GH and SPIDAR-GM are used for 3D tracking as well as for force rendering. Where SPIDAR-GH

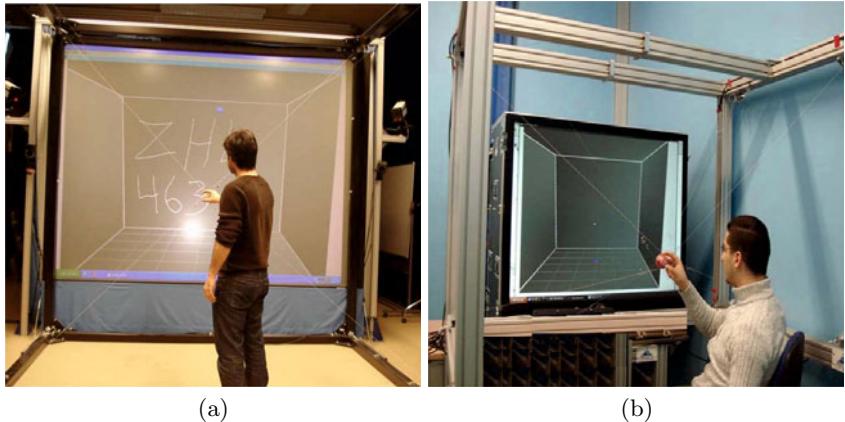


Fig. 3. Illustration of the dynamic haptic interaction in distributed CVE (a) User 1 with SPIDAR-GH of the semi-immersive platform (b) user 2 with SPIDAR-GM

and SPIDAR-GM means that both of them are SPIDAR-Gs (6DoF) but one is Human scale (H) and the other is Medium scale (M).

4 Evaluation

In this section we present the user study carried out to evaluate the effectiveness of our proposed guide.

4.1 Method

In order to evaluate the system, twenty volunteers participated in the experiment. They were master and PhD students and had ages from 22 to 30. Majority of them had no prior experience of the use of a haptic device. They all performed the experiment with the same expert. We installed our application on two machines (one for the expert and other for subjects) connected through local area network. Both machines were equipped with a phantom omni. Once the application was launched and the network connection successfully established between the two computers. Two spheres (red and blue) could be seen on both the screens. The red and blue spheres were assigned to the expert and subject respectively. The subjects were required to bring their sphere closed to the expert's sphere when the latter changes its color (i.e becomes green). In addition each subjects was asked that when the force starts guiding his/her hand, then he/she should not look to the screen, in order to avoid learning through visual movements of the cursor (sphere). On the termination of the guiding force he/she was required to write the alphabet or geometrical form on a paper to which he/she thinks his/her hand's movement correspond to. There were two guiding sessions. In the first session the subjects were guided to write ten alphabets (L, G, M, R, S, P, Z, W, B, N). In the second session they got guidance for four forms (triangle, square, circle and star). At the end they also responded to a questionnaire.

Table 1. User perception of the alphabets through haptic guide

Alphabets	L	G	M	R	S	P	Z	W	B	N
% of correct responses	100	90	100	85	95	90	85	80	100	100
Forms	Triangle	Square	Circle	Star						
% of correct responses	100	100	100	90						

Table 2. User perception of the Forms through haptic guide

Forms	Triangle	Square	Circle	Star
% of correct responses	100	100	100	90

4.2 Results Analysis

The table 1 depicts the percentage of correct and incorrect responses in case of guidance for alphabets. Here we see the overall percentage of correct response is high but still there are some subjects that did not perceive correctly some alphabets. This misperception was mainly due to the difference in the method of writing of a character by the expert and subjects. Similarly the table 2 shows that the first three geometrical forms were correctly responded by all subjects, only 10% subjects perceived the star incorrectly.

In response to the questionnaire, majority of the users were enthusiastic and reported that they found the experience very interesting. Similarly we asked them to mark the scale (1-2-3-4-5) for the level of guidance that they were provided. Here 1= small level of guidance and 5= high level of guidance. The average response was 3.84 (std: 0.68).

5 Conclusion

We present a new concept of “What You Feel Is What I Do” (WYFIWID). The concept is fundamentally based on a haptic guide that allows an expert to control the hand of a remote trainee. When haptic guide is active then all movements of the expert’s hand (via input device) in the 3D space are haptically reproduced by the trainee’s hand via a force feedback device. We used haptic guide to control the trainee’s hand for writing alphabets and drawing geometrical forms. Analysing the results of experiments we observed that the users found the haptic guidance very effective and interesting. This guide can be used to teach the writing of alphabets and digits of various languages and drawings to the children.

We plan to use and extend the concept of WYFIWID in the context of the DigitalOcean project to allow human-robot-interaction in mixed reality environments. In this case the user can feel the ROV (Remotely Operated underwater Vehicle) movements during the exploration mission.

Acknowledgements

This work is supported by the Conseil Général de l’Essonne, the IBISC laboratory and the DigitalOcean project.

References

1. Basdogan, C., Ho, C.-H., Srinivasan, M.A., Slater, M.: Virtual training for a manual assembly task. In: *Haptics-e*, vol. 2 (2001)
2. Chan, A., Maclean, K., Mcgrenere, J.: Designing haptic icons to support collaborative turn-taking. *International Journal of Human-Computer Studies* (2007)
3. David, M., Arnaldi, B., Plouzeau, N.: A general framework for cooperative manipulation in virtual environments. In: *Virtual Environments 1999 Proceedings of the Eurographics Workshop*, pp. 169–178 (1999)
4. Jordan, J., Mortensen, J., Oliveira, M., Slater, M., Tay, B.K., Kim, J., Srinivasan, M.A.: Collaboration in a mediated haptic environment. In: *The 5th Annual International Workshop on Presence* (2002)
5. Mortensen, J., Vinayagamoorthy, V., Slater, M., Steed, A., Lok, B., Whitton, M.C.: Collaboration in tele-immersive environments. In: *Proceedings of the Workshop on Virtual Environments*, pp. 93–101 (2002)
6. Otto, O., Roberts, D., Wolff, R.: A review on effective closely-coupled collaboration using immersive cve's. In: *VRCIA 2006: Proceedings of the 2006 ACM International Conference on Virtual Reality Continuum and its Applications*, pp. 145–154. ACM, New York (2006)
7. Plimmer, B., Crossan, A., Brewster, S.A., Blagojevic, R.: Collaborative virtual sculpting with haptic feedback. In: *Virtual Reality*, vol. 10, pp. 73–83. Springer, London (2006)
8. Plimmer, B., Crossan, A., Brewster, S.A., Blagojevic, R.: Multimodal collaborative handwriting training for visually-impaired people. In: *CHI 2008: Proceeding of the Twenty-sixth Annual SIGCHI Conference on Human Factors in Computing Systems*, pp. 393–402. ACM, New York (2008)
9. Prada, R., Payandeh, S.: On study of design and implementation of virtual fixtures. *Virtual Reality* 13(2), 117–129 (2009)
10. Riva, G., Bacchetta, M., Cesa, G., Molinari, E.: Cybertherapy: Internet and virtual reality as assessment and rehabilitation tools for clinical psychology and neuroscience, pp. 121–164. IOS Press, Amsterdam (2004)
11. Sallnas, E.-L., Rassmus-Grohn, K., Sjostrom, C.: Supporting presence in collaborative environments by haptic force feedback. *ACM Trans. Comput.-Hum. Interact.* 7(4), 461–476 (2000)
12. Otmane, S., Mallem, M., Kheddar, A., Chavand, F.: Active virtual guide as an apparatus for augmented reality based telemanipulation system on the internet, pp. 185–191. IEEE Computer Society, Los Alamitos (2000)
13. Ullah, S., Ouramdane, N., Otmane, S., Richard, P., Davesne, F., Mallem, M.: Augmenting 3d interactions with haptic guide in a large scale virtual environment. *The International Journal of Virtual Reality* 8, 25–30 (2009)
14. Ullah, S., Richard, P., Otmane, S., Mallem, M.: The Effect of audio and Visual aids on task performance in Distributed Collaborative Virtual Environments. In: *Intelligent Systems and automation: 2nd Mediterranean Conference on Intelligent Systems and Automation (CISA 2009)*. AIP Conference Proceedings, vol. 1107, pp. 196–201 (2009)
15. Oakley, I., Brewster, S., Gray, P.: Can You Feel the Force? An Investigation of Haptic Collaboration in Shared Editors. In: *Proceedings of Eurohaptics*, pp. 54–59 (2001)

16. Feygin, D., Keehner, M., Tendick, R.: Haptic guidance: experimental evaluation of a haptic training method for a perceptual motor skill. In: Proceedings of the 10th Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems, pp. 40–47 (2002)
17. Yoshikawa, T., Henmi, K.: Human skill transfer using haptic virtual reality technology. In: The 6th International Symposium on Experimental Robotics, vol. 250, pp. 351–360 (2000)
18. Dan, M., Hong, T., Federico, B., Timothy, C., Kenneth, S.: Haptic Feedback Enhances Force Skill Learning. In: Proceedings of the Second Joint EuroHaptics Conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems, pp. 21–26 (2007)
19. Bluteau, J., Gentaz, E., Coquillart, S., Payan, Y.: Haptic guidances increase the visuo-manual tracking of Japanese and Arabic letters. In: International Multisensory Research Forum (2008)
20. Chellali, A., Dumas, C., Milleville, I.: WYFIWIF: A Haptic Communication Paradigm For Collaborative Motor Skills Learning. In: Proceedings of the Web Virtual Reality and Three-Dimensional Worlds (2010)
21. Makoto, S.: Development of String-based Force Display: SPIDAR. In: IEEE VR Conference (2002)

A Framework Interweaving Tangible Objects, Surfaces and Spaces

Andy Wu, Jayraj Jog, Sam Mendenhall, and Ali Mazalek

Synaesthetic Media Lab,
GVU Center, Georgia Institute of Technology,
TSRB, 85 5th St. NW
Atlanta, GA 30318, USA
`{andywu, jayraj.jog, smendenhall13, mazalek}@gatech.edu`

Abstract. In this paper, we will introduce the ROSS framework, an integrated application development toolkit that extends across different tangible platforms such as multi-user interactive tabletop displays, full-body interaction spaces, RFID-tagged objects and smartphones with multiple sensors. We will discuss how the structure of the ROSS framework is designed to accommodate a broad range of tangible platform configurations and illustrate its use on several prototype applications for digital media content interaction within education and entertainment contexts.

Keywords: tangible interaction, API, interactive surfaces.

1 Introduction

In the past decade, researchers have begun to address the needs of developing interaction techniques that can more seamlessly bridge the physical and digital worlds. The vision of Tangible Media [1] presents interfaces that integrate physical objects with interactive surfaces and responsive spaces that embody digital information. Yet working with diverse sensing and display technologies is difficult and few tools support this type of development. Our research lab focuses on tangible interaction and sensing technologies that support creative expression bridging the physical and digital worlds. In the last few years, we have developed several projects based on a draft framework, the Responsive Objects, Surfaces and Spaces (ROSS), a unified programming framework that allows developers to easily build applications across different tangible platforms.

Tangible interaction research is a relatively young field, and as such application designers and developers working in this space face a number of challenges. The sheer number and variety of sensing techniques, interface devices, form factors and physical settings envisioned by researchers present a complex space from the perspective of generalizability of development tools. To further complicate the situation, many of the sensing and display technologies used by tangible interfaces are still in the process of evolution and maturation, and have not yet reached the state of stability and robustness.

2 Related Work

User interface toolkits enable programmers to rapidly create applications for graphical user interfaces (GUIs) that make use of traditional input/output devices. Tangible user interfaces combine non-standard hardware and software. A tangible toolkit allows developers to develop applications rapidly without having detailed knowledge of the underlying protocols and standards.

2.1 Multiouch Toolkits

Multitouch applications have become quite popular lately. There are many software toolkits that support multitouch gesture interaction and object manipulation. A lot of them share the same protocols, such as TUO or Open Sound Control (OSC). These toolkits mostly use a client-server architecture. In other words, these toolkits acts as servers that dispatch multitouch messages to client applications.

libTISCH [2] (Library for Tangible Interactive Surfaces for Collaboration between Humans), is a development framework for tabletop computers. It provides a number of features such as gesture recognition, GUI widgets and hardware drivers for various tangible interaction input devices and technologies (FTIR, DI, Nintendo Wii, Microsoft Kinect). libTISCH employs a layered architecture that consists of four layers: hardware abstraction, transformation, interpretation and widget. The hardware abstraction layer takes raw data from the input devices regarding the positions of objects, finger touches etc. and produces data packets. The transformation layer converts the raw data into corresponding screen coordinates, which then cause the interpretation layer to generate events as appropriate. The widget layer listens for events and reacts accordingly.

ReacTIVision [3] is a computer-vision based system designed for interactive tabletop displays. It tracks not only finger touches but also objects tagged with computer vision patterns, called fiducial markers. ReacTIVision defines its own tangible interaction protocol, the TUO protocol to communicate with client applications. The TUO is an open framework based on Open Sound Control – an emerging standard for interactive environments. The TUO protocol and its API have become a standard for the interactive surface community. Several of the early ROSS applications adopt the TUO protocol. Another similar project, TWING is a software framework that has been used to develop the Xenakis table [4], a multi-user interactive tabletop display to create music through objects and finger gestures.

2.2 Physical Device Toolkits

Physical device toolkits often provide predesigned hardware components and software APIs for rapid prototyping. Phidgets [5] are electronic components analogous to software widgets, allowing construction of complex physical systems out of simpler components. Phidgets provide the physical components and the software API that lower the barrier for novices to control electronic devices through computers. Similar to Phidgets, the Microsoft Gadgeteer [6] is a prototyping platform for rapid constructing and programming. However, it also integrates 3D CAD support for developers to design the physical container for the electronics. Introducing a CAD

library to the toolkits helps developers build embedded system products even more efficiently. Toolkits for developing Augmented Reality (AR) or Virtual Reality (VR) applications have the capability of detecting objects in the 3D space. They are not toolkits for physical devices but still worth mentioning here. The ARToolkit [7] is a computer vision based system that tracks fiducial markers attached to objects. Another toolkit, DART [8] is designed for rapid development of AR or VR applications through GUI.

2.3 Integrated Toolkits

Prior work on tangible application development has shown that both these approaches are important in evolving the area of TUI design, and we thus feel they should not remain entirely separate from one another. In order to further support these developments, the toolkits for shared tangible display surfaces need to support communication with other kinds of devices as well.

The iROS platform [9] supports an interactive workspace that incorporates technologies on many different scales, from personal devices to large table and wall displays. The iROS utilizes a dedicated server to dispatch customizable messages between different devices. The Papier-Mâché toolkit [10] provides an event-based model for application development using RFID-tagged, barcoded, or computer vision identified objects. The empirical study of Papier-Mâché also discovered several necessary features that influenced our tangible toolkit design, which are ease of use, facilitating reuse and modular code structure.

Our previous work on the Synlab API [11] has integrated different types of tangible platforms into a shared framework. The Synlab API is based on the API of the TVViews [12] media table, a multi-user interactive platform that supports real-time tracking of tagged objects. We have extended the Synlab API to include additional technologies and platforms, including RFID object tracking, multi-touch surfaces, and displays such as a tilting table and spinning screen.

TwinSpace is a generic framework for developing collaborative cross-reality environments. It allows the connection of heterogeneous clients (mouse + keyboard clients, interactive tabletop display, instrumented spaces, etc.) to a shared virtual environment. The TwinSpace architecture is presented in detail in [13]. As a platform for prototyping and experimentation, the TwinSpace infrastructure permits a range of techniques for integrating physical and virtual spaces. Events generated by sensors or input devices in a physical space can be published and reported in the virtual space. Furthermore, TwinSpace offers facilities to define spatial and structural correspondence between physical and virtual spaces (for example, the meeting space can be mapped to a corresponding region in the virtual space). Spaces may also be connected through functional ‘contact points’ whether the real and virtual spaces correspond spatially or not (for example, a tabletop interface might present an interactive top-down view over the virtual space).

3 ROSS Infrastructure

Working with emerging sensing and display technologies is a challenging task for application developers. In order to integrate heterogeneous tangible interfaces we

provide the ROSS programming framework that is based on a nested abstraction of responsive objects, surfaces and spaces, for developing tangible user interfaces. This serves to unify the application development process across a broad range of the TUI space, and can accommodate the continued evolution of the underlying technologies and devices.

3.1 Relationship between Objects, Surfaces and Spaces

The relationships between the three cornerstones of ROSS, objects, surfaces and spaces, can be understood simply by talking about their respective roles. A ROSS object is any device that is capable of running a ROSS application. Thus, peripheral devices such as cameras, microphones, mice etc. are not ROSS objects, while computers, mobile phones, tablets, tables and microcontrollers are ROSS objects. A ROSS surface usually refers to any 2-dimensional space with which it is possible to interact. This could mean touchscreens, touch-sensitive walls, tabletop computers or any other surface through which it is possible to provide input to a ROSS application. ROSS objects can be placed on ROSS surfaces.

A ROSS space is any 3-dimensional space in which one or more ROSS objects or surfaces are located and within which it is possible to interact with a ROSS application. ROSS spaces abstract the various problems of locating a ROSS object spatially, in addition to being a means of input through gestures, sounds, light patterns etc.

3.2 ROSS Design Goals

Tangible tools need to share common functionality across different applications, provide standards that can support development across different kinds of platforms and devices, and most important of all is the ease of use. ROSS has the following design goals:

- Designed for applications. The design of ROSS API is through a practice-based approach. In other words, we have built and continue to build applications based on the concept of ROSS and improve the API from the developers' feedback. This means that ROSS API should be appropriate for a wide variety of contexts, applications and platforms.
- Platform independence. Although ROSS API is currently available in Java only, we aim to ensure that it will eventually be available in other popular programming languages. ROSS API introduced ROSS XML, a file that can be easily processed by most programming languages.
- Rapid development and scalable structure. ROSS is envisioned to work in a rapidly changing and evolving environment. It is also designed for critical network conditions. ROSS passes only the most essential messages for real-time communication. Using UDP for the transport layer achieves all of these goals. Besides, there is little connection maintenance overhead compared to TCP, and the lack of flow control and acknowledgment are ideal for real-time data transfer. In addition, ROSS can scale well without maintaining persistent connections.
- Work under resource constraints. Many of the applications for ROSS involve one or more embedded systems, e.g. mobile phones or single chip microcontrollers. Since these devices have memory and/or energy constraints, ROSS is designed with these constraints.

3.3 ROSS Structure

The core of the ROSS framework rests on the notion that the different objects, surfaces and spaces in a responsive environment can exist in nested relationships with respect to one another. The concept of nested objects, surfaces and spaces is most easily illustrated using an example, as follows.

Consider an RPG game that runs on an interactive table. Users interact with the game through tagged physical pawns and finger touches on the sensing surface. We can first think of the interactive table itself as a responsive object that provides a nested sensing surface. Each tagged pawn is another kind of interactive object that is in a nested relationship to the sensing surface. Now imagine we replaced the pawns with android phones. These objects now each provide their own nested sensing surface, on which finger touches can again be seen as nested objects. The phone might also provide interaction through touches, buttons or other sensors, which can be considered as nested controls that operate on the interactive object.

ROSS devices have an XML file that provides more information about those devices. A ROSS device can be just a ROSS object or also have a ROSS surface or be in a ROSS space.

Table 1. A sample ROSS XML structure. The server-side XML indicates that the server table needs to listen to events from *Android* phones. The client-side XML indicates that the client phone needs to listen to events from a *TuioTable*.

Server-side XML (table)	Client-side XML (Android phone)
<pre><Device type="TuioTable" id="0"> </Device> <Subscription type="Android"> </Subscription></pre>	<pre><Device type="TuioTable" id="0"> </Device> <Subscription type="Android"> </Subscription></pre>

As you can see, the ROSS XML file describes the structure of the ROSS device to which it belongs. The XML encapsulates the nested structure of the device perfectly. This particular device (a cellphone) is a ROSS object that also has a ROSS surface. Other devices (such as cameras) have the ability to understand ROSS space and their relationship to it.

3.4 Cross Network Communication

The ROSS API is designed as a distributed networked system. Each node of a ROSS network acts as a client, receiving information about other I/O devices, and also as a server, distributing information about the I/O devices that it contains.

Each node contains XML file that holds a description of its I/O devices. By trading this information with other nodes, each node knows the device capabilities of the other nodes in the network. This file also contains information on what devices the particular node wants to listen to. When a node parses a remote XML file, it looks

through this information and subscribes the remote node to the I/O events of any local matching devices. The I/O events or state of a particular device are transmitted to subscribers via a UDP protocol. A ROSS node both listens to, and broadcasts its IP address from a multicast group. Upon receiving the multicast packet, another ROSS node will establish a TCP/IP connection with the packet sender. The nodes then exchange their ROSS XML files.

Due to the modular driver/XML parsing architecture, to construct a new ROSS application one only needs to gather all the necessary device representation objects and drivers, write an appropriate XML device structure and have the application listen for new devices and register for their I/O events. No networking code is needed for sending and receiving I/O events. The ROSS Manager handles it.

4 Applications

We have shown that it is possible to specify a broad range of different platforms using the ROSS structure. We describe some selected example applications below.

4.1 Moons Over You

Moons Over You [14] is one of the earliest projects that experimented with the ROSS infrastructure. In the project room, a moon appears projected on the wall for each person entering the room, and each person's moon then follows them around as long as they remain in the space. At the same time, these moons can be seen on the surface of the interactive tabletop display. Users at the table can visualize people's movements by looking at the changing image of a moon-filled sky on the table's surface. They can also reach into this tabletop sky, and playfully re-arrange the moons with their finger touches. The moons respond to these finger touches as well, and people thus get the sense that their moon is playing tricks on them as it runs away. The left figure of Fig. 1 illustrates how the server captures and interprets the movement of users and sends the ROSS messages to a client dedicated for visualizing the actions through the wane and wax of moons. The right photo of Fig. 1 shows several moons based on the users' locations and movement.

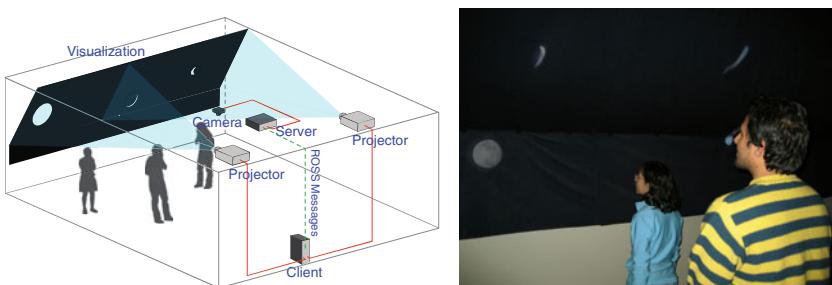


Fig. 1. The system diagram of Moons Over You and the users' interaction

4.2 ROSS Ripple

ROSS Ripple is an initial concept and technology demo built with the ROSS API. It is an exploration of embedded and synchronized interactive surfaces. Ripple connects an interactive tabletop and Android phones with fiducial markers attached to their backs. When the user touches any of the surfaces, ripples of color emanate from the touch position. When placed on the table, a phone appears to become part of the table surface. The backgrounds of the table and phone align. Ripples created on the table surface run across phones. Ripples created on the phone surfaces run onto the table. If a phone is lifted off of the table, then touches on the phone have a one-to-one “synchronized” spatial relationship with ripples on the table surface.

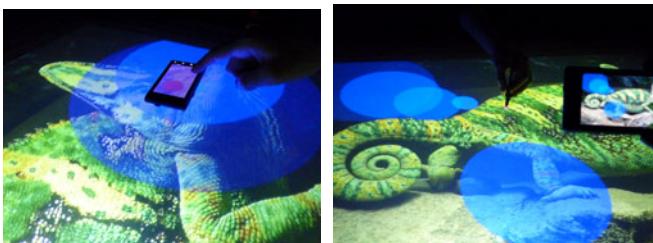


Fig. 2. The interactions of ROSS Ripple

In the left photo of Fig. 2, a mobile phone is placed on an interactive tabletop display. The blue ripples on the tabletop are generated by touches on the tabletop while the purple ripple on the cellphone screen is generated by a touch on the cellphone screen. These two types of ripples are interwoven because of the nested structure of ROSS. The screen of mobile phone acts like a small window on the tabletop that reveals the tabletop image blocked by the phone. In the right photo of Fig. 2, the mobile phone has left the tabletop surface. The cellphone screen and the tabletop display show exactly the same image. The act of placing one surface onto another embeds that surface into the surface beneath.

4.3 Ross Paint

ROSS Paint is a sketching application that utilizes multiple phones as touch input devices to paint onto a monitor or screen. Lines are drawn between touch inputs from each phone to produce a unique line drawing effect. ROSS Paint explores how to integrate multiple input devices into collaborative artwork. Users do not merely interact separately on the same surface. The nature of the tool forces them to synchronize their activities to produce a desired effect. Fig. 3 shows the pattern when a user draws random curves on two phones.

4.4 Kinect Applications

We propose a class of applications that employ the Kinect sensor manufactured by Microsoft. The Kinect allows reconstructing the 3D scene of a space. It consists of IR

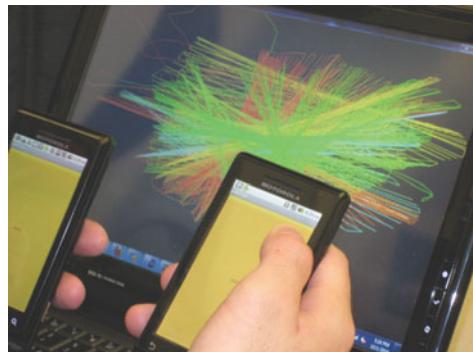


Fig. 3. ROSS paint with two cellphones

and visible light cameras and microphones, and can be combined with sophisticated image recognition software to detect fiducial markers, objects, and even perform facial recognition. A ROSS space can be defined in terms of the Kinect's location in it. It can act as the “origin” of the coordinate system and the other devices can locate themselves with respect to each other in this system. Already networked Kinects are used in pairs to build 3D models of a room. They may be used to perform perfect object recognition and thus obviate the need for fiducial markers.

Kinect-based ROSS application – Impromptu tangible objects. Kinect is used to recognize shapes of objects. This, combined with the in-built microphone can be used to “teach” a table to recognize objects as tangibles. By further associating them with pre-existing objects whose properties are stored in a database, it is possible to use an arbitrary object as a proxy for any other object, for the sake of convenience (as shown in Fig. 4). For example, if a user wants to discuss the possible positioning of various types of billboards on a network of highways with some associates, and all she brought are pens, pencils and notepads. She could conceivably train the table to recognize each pen or pencil as a different type of billboard. When a pen is placed on the table, the table could display an image of the corresponding billboard next to the pen.

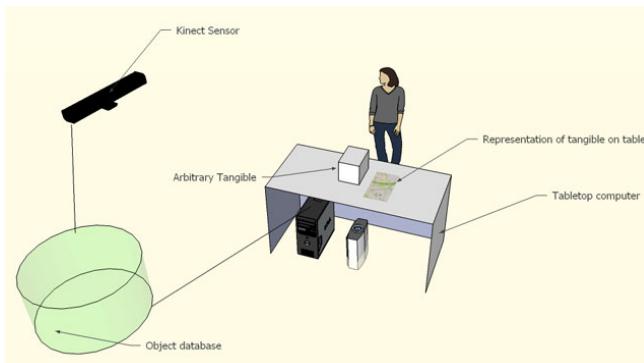


Fig. 4. The concept diagram of ROSS Kinect

5 Summary

This paper has demonstrated the capability of the ROSS API through several applications. In order to test out and develop a breadth of applications and interaction concepts on emerging tangible platforms, creators need easy development tools to work with. The ROSS API has integrated heterogeneous resources, from multiple objects and surfaces to spaces. In the future, we plan to continue extending the ROSS API to support a broader range of platforms and technologies, and to support application development in different programming languages. The current version of the ROSS API is an engineering prototype. Working with emerging sensing and display technologies is a challenging task for application developers. In order to support continued growth of tangible interfaces, we need to improve the ROSS API from an engineers' prototype to an easy-to-use programming tool.

Acknowledgments. We thank all the application developers who used and provided feedback on ROSS in its early stages.

References

1. Ishii, H., Ullmer, B.: Tangible bits: towards seamless interfaces between people, bits and atoms. In: Pemberton, S. (ed.) *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 234–241. ACM Press, New York (1997)
2. Echtler, F.: libTISCH: Library for Tangible Interactive Surfaces for Collaboration between Humans, <http://tisch.sourceforge.net/>
3. Kaltenbrunner, M., Bencina, R.: reactIVision: a computer-vision framework for table-based tangible interaction. In: *Proceedings of the 1st international conference on Tangible and embedded interaction*, pp. 69–74. ACM Press, New York (2007)
4. Bischof, M., Conradi, B., Lachenmaier, P., Linde, K., Meier, M., Pötzl, P., André, E.: Xenakis: combining tangible interaction with probability-based musical composition. In: *Proceedings of the 2nd international conference on Tangible and embedded interaction*, pp. 121–124. ACM Press, New York (2008)
5. Greenberg, S., Fitchett, C.: Phidgets: easy development of physical interfaces through physical widgets. In: *Proceedings of the 14th annual ACM symposium on User interface software and technology*, pp. 209–218. ACM Press, New York (2001)
6. Villar, N., Scott, J., Hodges, S.: Prototyping with Microsoft.NET Gadgeteer. In: *Proceedings of TEI 2011*, pp. 377–380. ACM Press, New York (2011)
7. Kato, H., Billinghurst, M., Poupyrev, I., Imamoto, K., Tachibana, K.: Virtual object manipulation on a table-top AR environment. In: *International Symposium on Augmented Reality, IEEE and ACM International Symposium on Augmented Reality*, pp. 111–119. IEEE Press, New York (2000)
8. MacIntyre, B., Gandy, M., Dow, S., Bolter, J.D.: DART: a toolkit for rapid design exploration of augmented reality experiences. In: *Proceedings of the 17th annual ACM symposium on User interface software and technology*, pp. 197–206. ACM Press, New York (2004)
9. Ponnekanti, S.R., Johanson, B., Kiciman, E., Fox, A.: Portability, Extensibility and Robustness in iROS. In: *Proc. First IEEE international Conference on Pervasive Computing and Communications*, pp. 11–19. IEEE Press, New York (2003)

10. Klemmer, S.R., Li, J., Lin, J., Landay, J.A.: Papier-Mâché: toolkit support for tangible input. In: Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 399–406. ACM Press, New York (2004)
11. Mazalek, A.: Tangible Toolkits: Integrating Application Development across Diverse Multi-User and Tangible Interaction Platforms. In: Let's Get Physical Workshop, 2nd International Conference on Design Computing and Cognition, July 2006, Eindhoven University of Technology, Netherlands (2006)
12. Mazalek, A., Reynolds, M., Davenport, G.: TViews: An Extensible Architecture for Multiuser Digital Media Tables. *IEEE Computer Graphics and Applications* 26(5), 47–55 (2006)
13. Reilly, D.F., Rouzati, H., Wu, A., Hwang, J.Y., Brudvik, J., Edwards, W.K.: TwinSpace: an infrastructure for cross-reality team spaces. In: Proceedings of the 23rd annual ACM symposium on User interface software and technology, pp. 119–128. ACM Press, New York (2010)
14. Lee, H.J., Wu, C.S.(A.), Shen, Y.T., Mazalek, A.: Moons over you: the poetic space of virtual and real. In: ACM SIGGRAPH 2008 posters, Article 24, 1 pages. ACM Press, New York (2008)

The Effect of Haptic Cues on Working Memory in 3D Menu Selection

Takehiko Yamaguchi, Damien Chamaret, and Paul Richard

Laboratoire d'Ingénierie des Systèmes Automatisés (LISA),

Université d'Angers, 62 Avenue ND du Lac – 49000 Angers, France

{takehiko.yamaguchi, damien.chamaret, paul.richard}@univ-angers.fr

Abstract. We investigated the effect of haptic cues on working memory in 3D menu selection. We conducted a 3D menu selection task in two different conditions: visual only and visual with haptic. For the visual condition, participants were instructed to select 3D menu items and memorize the order of selection. For the visual with haptic condition, we used magnetic haptic effect on each 3D menu item. Results showed that participants needed less number of trials for memorizing the selection sequence in visual with haptic condition than in visual only condition. Subjective data, collected from a questionnaire, indicated that visual with haptic condition was more suitable for selection and memorization.

Keywords: Virtual reality, haptic interaction, 3D menu, selection, learning.

1 Introduction

Humans acquire new motor skills through a multi-stage learning process, central to which is of course practice. Through the process of trial and error, we continually refine our motor skills achieving more consistent performance. In this context, feedback is of paramount importance and can take many different forms including verbal communication, visual and auditory signals, and vibrotactile stimulation. Although different in sensory modality, these examples all share the common trait of being indirect forms of feedback. That is, the information that they provide about performance must be translated into the proprioceptive coordinate system. For simple tasks, this translation may not be a significant difficulty, but for more complex tasks it may be overwhelming, particularly in the early stages of learning. More direct forms of feedback are physical guidance where the learner is physically moved all along a trajectory or where the learner is closed to the target.

Current technology makes use of motion based interfaces such as WiimoteTM or KinectTM. This trend indicates that next generation of human interfaces will involve 3D interactions and complex tasks such as 3D menu selection that require efficient learning and memorization. 3D space is recognized in obtaining information from a variety of visual depth. However, it is difficult to select the 3D menu items in 3D virtual space with only visual information so that other modality information such as haptic or audio cue has been used to improve selectability of the 3D menu items.

Generally, an application interface has a well-categorized menu structure. When a user calls an integrated function on the application, a couple of menu items are needed to select. In the meantime, the sequence of the selection process should be memorized to learn about how to use the application. Since these processes: selection and memorization are usually carried out at the same time, we could expect that the operability of the selection affects the memorization of the menu selection process. In fact, it is important to investigate the effect of the selectability to address workload of the memorization. However most of studies have focused on operation performance.

In this study, we investigated the effect of haptic cues on working memory in 3D menu selection. We conducted a 3D menu selection task in two different conditions: visual only and visual with haptic. For the visual condition, participants were instructed to select 3D menu items and memorize the order of selection. For the visual with haptic condition, we used magnetic haptic effect on each 3D menu item. In the next section, we present some previous works. In section 3, we describe our system. Section 4 is dedicated to the experiment. The paper ends by a conclusion that provides some tracks for future work.

2 Background

A number of studies have shown that haptic feedback can improve user performance in various tasks involving cursor control [1-4]. In particular, force feedback gravity wells, i.e. attractive basins that pull the mouse and cursor to the center of a target, have been shown to improve performance in “point-and-click” tasks. Hasser et al [5] found that this type of force feedback, provided by a FEELit mouse, could improve targeting time and decrease errors. Oakley et al [6] reported a reduction in errors with the use of gravity wells implemented on a Phantom. Keates et al [7] found that for motion-impaired users, gravity wells could improve the time required to complete a “point-and-click” task by as much as 50%. In these studies, however, force feedback was enabled on a single target only. For the successful implementation of force feedback in a realistic interface, issues surrounding haptic effects for multiple on-screen targets must be addressed. With more than one gravity well enabled, a user’s cursor may be captured by the gravity wells of undesired distractors as it travels toward a desired target.

This has the potential to cancel out the benefits of the force feedback, possibly yielding poorer performance than in its complete absence. There have been few studies investigating performance in the presence of multiple haptic targets. Dennerlein and Yang [8] found that even with multiple haptic distractors along the cursor trajectory, performance in “point-and-click” tasks was greatly improved over a condition with only visual feedback. Study participants most often just plowed through the distractors, but at a cost of increased user frustration and effort. In contrast, Oakley et al [9] reported an increase in time when static attractive forces were enabled on multiple targets. This condition was, at best, not optimal, and at worst, detrimental to performance and subjective satisfaction when compared to the purely visual condition. Langdon et al [10] reported a performance improvement for motion-impaired users that was similar across four sizes of gravity wells on adjacent targets.

Physical guidance refers to the process of physically moving a learner's body through an example of a motion to be learned. Haptically enhanced interaction for guidance mainly relies on "snap-to" effects. They can be local magnetic effects around a target that actively captures the pointer if it enters a specific area, or can behave as a gradient force all over the environment to draw the pointer towards points of interest. For object selection, magnetic targets can help by reducing selection times and error rates. However some studies report benefits from magnetic widgets to precision but not to selection times [12]. Moreover, these techniques seem to lead to higher selection times and to a significantly higher overall cognitive load when multi-target selection is considered. Capobianco and Essert [13] proposed a technique able to reduce these drawbacks, and apply it in the context of item selection in 3D menus. Their approach called "haptic walls" consists in haptically rendering solid walls shaped like a funnel, leading to a target located at the intersection of the two walls. The walls act as virtual fixtures: the targets are accessible while slipping along the interior faces and edges of the convex polyhedron that connects them. This approach differs from a magnetic grid since the edges of the haptic shape are not attracting the pointer towards them. This technique can be adapted to any configuration of targets able to be represented as a convex polyhedron. Yamaguchi et al [14]., investigated the usability of haptic interface features to support future development of haptically enhanced science learning applications for students with visual impairments. Three features have been proposed in this study: Haptic Boundary, Menu Selection, and Object Recognition. Results gave insights to the needs of the visually impaired community in regard to haptic applications.

As described above, most of study has focused on investigating capability of haptic cues, especially for task performance which enables to improve completion time or decrease error rate, and so on. Han et al [11]., reported the design and results of an experiment that evaluated the effects of enactment and haptic guidance on building a memory of a sequence of 2D selections, with learning only with visual information as a baseline condition. As an initial study, they focused on the role of the working memory rather than that of the long term memory, since the working memory chunking takes precedence over building the long term memory. For a training system, a large manipulator-type haptic interface was combined with a large LCD visual display to cover the whole arm workspace involved in the task, realizing a collocated visuo-haptic training system. To test the effects of the kinesthetic information, they compared the learning performances of three training methods: visual display only, visual display with enactment, and visual display with haptic guidance. The learning performances were measured in terms of the number of repetitions required for memorizing, along with subjective ratings. However, they has not investigated the effect of haptic cues on menu item selection task in 3D virtual environment.

3 System Overview

The system we use for the study consists of a visual display, a 6 DOF haptic display called SPIDAR-GCC [15], and a Desktop computer. Fig.1 shows the hardware

configuration of the system. SPIDAR-GCC is connected to the Desktop PC through USB 2.0 communication. The grip of SPIDAR-GCC is associated with virtual pointer which is a visual pointer displayed in the visual display for operation. Since the workspace of the grip of the SPIDAR-GCC is limited in the frame and does not have clutch function to put a grip position into neutral, the grip position is controlled with velocity control. As for the display, we use a WSXGA LCD monitor with resolution of 1680x1050 pixels.

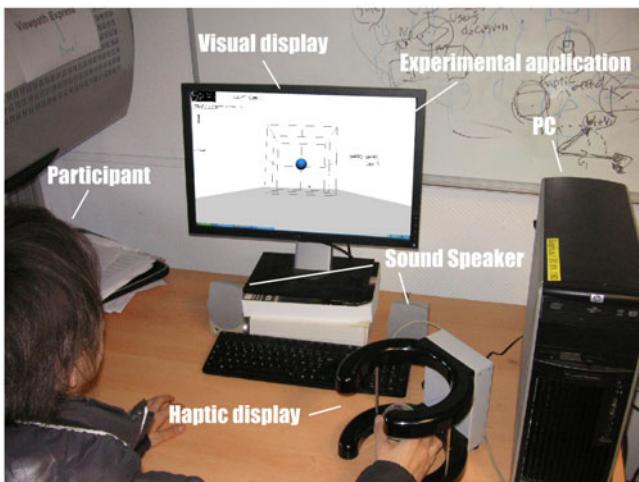


Fig. 1. Hardware configuration of the system - As for the haptic display SPIDAR-GCC, there is a grip which is connected in center of frame using 8 strings. This grip enables to input 6DOF information and represent 6DOF force feedback.

4 Method

In this study, we tested 3D menu selection task with two different conditions: visual condition, and visual with haptic condition to investigate whether difficulty of selection affects memorization of the selection order. In this task, cube type of menu is displayed on 3D grid (Fig.2). The participant selects the displayed menu one by one, and at the same time the participant has to memorize the order of selected cubes. We expected that the number of trial which is spent to memorize the order depends on the difficulty of selection. So, we expected if there is no stress on the selection, the number of trial is reduced.

4.1 Participants

Twenty-four university students from 20 to 24 years old participated in the experiment as participants. All of participants were male and were undergraduate students in author's university.

4.2 Stimuli

Visual condition (V). In total 9 blue cubes out of 12 cubes are randomly displayed on gray 3D grid one by one every 2 seconds as shown in Fig.2. The number in Fig.2 illustrates display order of each blue cube. Every time the total number of the displayed cubes is three, the displayed cubes are erased to refresh the screen. When touching the displayed cube with 3D pointer, sound effect is played as a notification for touching.

Visual with Haptic condition (VH). 9 blue cubes are displayed on 3D grid as same as the visual condition. When touching the cube with 3D pointer, magnet force is represented so that the 3D pointer is immediately put on the target cube. The amount of magnet force is calculated for ease of selection and release.

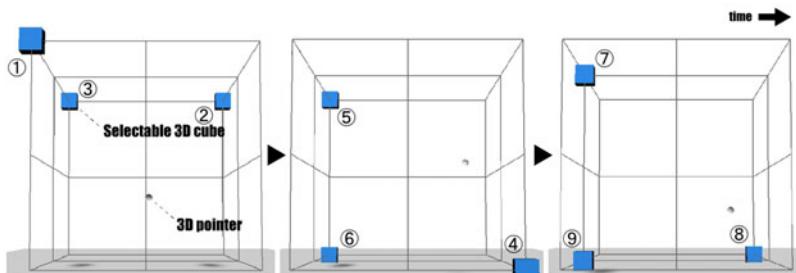


Fig. 2. 9 cubes are randomly displayed one by one in one trial

4.3 Procedure

We tested two conditions as explained in section 4.2. In both condition, same procedure was conducted.

Step1: When the experiment application is launched, the start screen as shown in Fig.3 is displayed. A Participant grabs the grip of SPIDAR-GCC and moves the 3D pointer to the blue sphere which is displayed in center of 3D grid. The position of the 3D pointer is fixed with magnet force.

Step2: The participant pushed the space button on the keyboard to begin a trial. The participant has to select displaying cube on the 3D grid one by one and at the same time, the order of cube has to be memorized correctly. When 9 cubes are displayed in total, the trial is finished.

Step3: If the participant is for sure that the order is perfectly memorized, answering mode is conducted. If not, they could try the trial again up to 20 times. In the answering mode, all 12 cubes are displayed with 10% transparency rate as shown in Fig.4. While touching the cube, the transparency rate is changed to 100%. The participant has to select the cube one by one to recall the memorized order and space button on the keyboard is pushed to verify the order. The limitation time of answering is set up to 90 seconds.

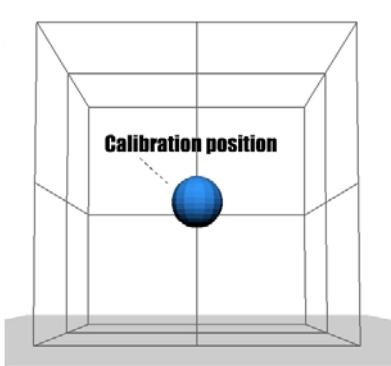


Fig. 3. Start page of experiment application

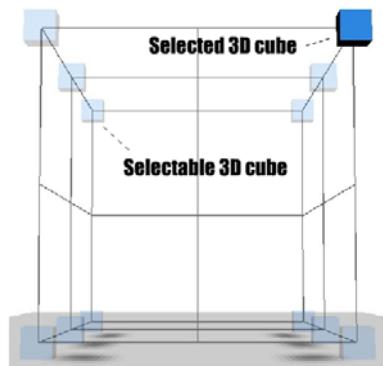


Fig. 4. Answering mode

In this experiment, we use between-subjects design to alleviate the affect of learning effect of each experiment condition. 12 participants were assigned to each condition to make two experiment groups.

5 Results and Discussion

5.1 Repetition Number of Trial

Fig.5 shows the results of average of repetition number of trial. A t-test was performed on the mean number of both conditions. T-test showed that there is significant difference between visual condition and visual with haptic condition (** p < 0.01). The result shows that participants were able to memorize menu on visual with haptic condition much more quickly than visual one.

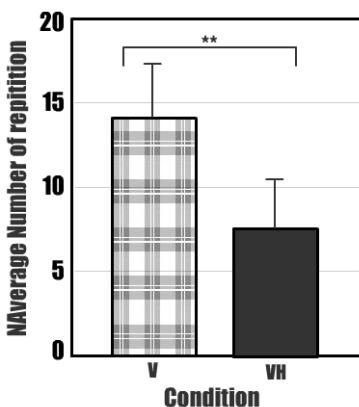


Fig. 5. Average of repetition number of trial

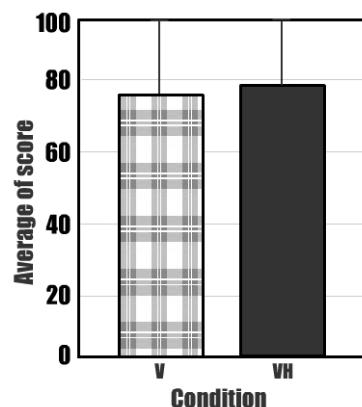


Fig. 6. Average of score of trial

5.2 Score of Trial

Fig.6 shows the results of average of score of trial. A t-test was performed on the mean number of both conditions. T-test showed that there is no significant difference between visual condition and visual with haptic condition. However, the result shows that there is a trend where score on visual with haptic is higher than the one on visual condition since the mean value on visual with haptic condition is higher than the one of visual condition. From this result and the result of repetition number, with the visual with haptic condition, the participants were able to memorize 3D menu order much more correctly and quickly than on the visual one.

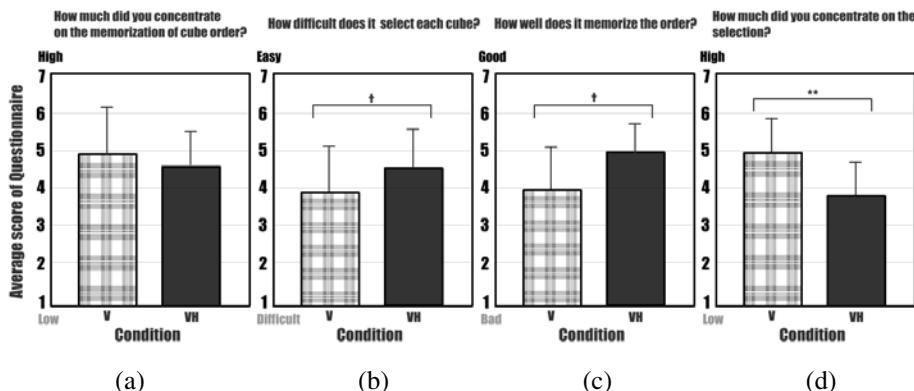


Fig. 7. Average score of questionnaire: (a) – Question 1, (b) – Question 2, (c) – Question 3, and (d) – Question 4

We asked 4 questions in a questionnaire after the trial session: **Question1** - *How much did you concentrate on the memorization of cube order?*, **Question2** - *How difficult does it select each cube?*, **Question3** - *How well does it memorize the order?*, and **Question4** - *How much did you concentrate on the selection?* A t-test was performed on the mean value of score of each question.

5.3 Concentration on the Memorization of Menu Order (Question 1)

The result shows that there is no significant difference and indicated participants were able to concentrate on the memorization on each condition. However, there is a trend where participants were able to concentrate on the memorization on visual condition much better than on the visual with haptic condition. From the result, we expected that visual attention might be on top of haptic attention. In other words, these attention functionalities might be independently working.

5.4 Difficulty on the Selection (Question 2)

The result shows that there is no significant difference, however there is significant trend so that the result indicated participants were able to select 3D menu easily on

visual with haptic condition. From the result, we expected if there is no difficulty on selection, participants would be able to concentrate on memorization since selection and memorization were done at the same time.

5.5 Difficulty on the Memorization (Question 3)

The result shows that there is no significant difference. However there is significant trend where participants were able to memorize 3D menu order on visual with haptic condition much better than on visual condition.

5.6 Concentration on the Selection (Question 4)

The result shows that there is significant difference between visual condition and visual with haptic condition (** $p < 0.01$). This result indicated participants had not paid attention to selection on visual with haptic condition. This result implies that participants could concentrate on memorization since the participant's attention was not used a lot for selection.

6 Conclusion

In this paper, we investigated the effects of haptic cues on memorizing the selection sequences in 3D menu. We tested it on two conditions: visual condition, and visual with haptic condition. As for the visual condition, participants selected 3D menu on 3D grid with 3D pointer and at the same time they memorized the order. For the visual with haptic condition, we controlled difficulty of selection by using haptic modality to put a magnetic haptic effect on each 3D menu. As the result of performance data, repetition number of trial where is a number the participants spent to memorize the order on visual with haptic condition is shorter than on visual condition. Moreover there is a trend where the score of trial on visual with haptic condition is better than visual condition. As a result from the subjective data, we expected visual attention is working on top of haptic attention since there is a trend where concentration on memorization on visual condition was higher than visual with haptic condition although they could memorize the order on visual with haptic with more fewer repetition number of trial than on visual condition (Question 1). As for the difficulty of selection, the participants were able to select 3D menu easily on visual with haptic condition (Question 2). For the difficulty of memorization, the participants were able to memorize the order easily on visual with haptic condition (Question 3). Moreover as for the concentration of selection, the participants were able to select easily on visual with haptic condition and this result indicated that they could memorize selection order very well (Question 4). In the future, we plan to investigate the effect of different modality and difficulty condition to address workload of memorization of 3D menu sequence.

Acknowledgments. This work was supported in the framework of the ReV-TV project by the French Ministry for Industry (DGCIS).

References

1. Feygin, D., Keehner, M., Tendick, F.: Haptic guidance: Experimental evaluation of a haptic training method for a perceptual motor skill. In: Proceedings of the Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems, pp. 40–47 (2002)
2. Oakley, I., Brewster, S., Gray, P.: Solving multi-target haptic problems in menu interaction. In: Proceedings of ACM CHI 2001: extended abstracts, pp. 357–358 (2001)
3. Wall, S.A., Paynter, K., Shillito, A.M., Wright, M., Scali, S.: The Effect of Haptic Feedback and Stereo Graphics in a 3D Target Acquisition Task. In: Proc. Eurohaptics 2002, pp. 23–29 (2002)
4. Akamatsu, M., MacKenzie, I.S., Hasbrouq, T.: A comparison of tactile, auditory, and visual feedback in a pointing task using a mouse-type device. *Ergonomics* 38, 816–827 (1995)
5. Hasser, C., Goldenberg: A User performance in a GUI pointing task with a low-cost force-feedback computer mouse. In: Proceedings of the ASME Dynamic Systems and Control Division, American Society of Mechanical Engineers, pp. 151–156 (1998)
6. Oakley, I., McGee, M.R., Brewster, S.A., Gray, P.D.: Putting the feel in look and feel. In: Proceedings of CHI 2000, pp. 415–422. ACM Press, New York (2000)
7. Keates, S., Hwang, F., Langdon, P., Clarkson, P.J., Robinson, P.: Cursor measures for motion-impaired computer users. In: Proceedings of ASSETS 2002, pp. 135–142. ACM Press, Edinburgh (2002)
8. Dennerlein, J.T., Yang, M.C.: Haptic force-feedback devices for the office computer: performance and musculoskeletal loading issues. *Human Factors* 43(2), 278–286 (2001)
9. Oakley, I., Adams, A., Brewster, S.A., Gray, P.D.: Guidelines for the design of haptic widgets. In: Proceedings of BCS HCI 2002, pp. 195–212. Springer, London (2002)
10. Langdon, P., Hwang, F., Keates, S., Clarkson, P.J., Robinson, P.: Investigating haptic assistive interfaces for motion-impaired users: Force-channels and competitive attractive-basins. In: Proceedings of Eurohaptics 2002, Edinburgh UK, pp. 122–127 (2002)
11. Han, G., Lee, J., Lee, I., Jeon, S., Choi, S.: Effects of Kinesthetic Information on Working Memory for 2D Sequential Selection Task. In: IEEE Haptics Symposium 2010, USA (2010)
12. Oakley, I., Brewster, S., Gray, P.: Solving multi-target haptic problems in menu interaction. In: Proceedings of ACM CHI 2001: extended abstracts, pp. 357–358 (2001)
13. Capobianco, A., Essert, C.: Study of Performances of “Haptic Walls” Modalities for a 3D Menu. In: Kappers, A.M.L., van Erp, J.B.F., Bergmann Tiest, W.M., van der Helm, F.C.T. (eds.) EuroHaptics 2010. LNCS, vol. 6192, pp. 152–159. Springer, Heidelberg (2010)
14. Yamaguchi, T., Johnson, S., Kim, H.N., Li, Y., Nam, C.S., Smith-Jackson, T.L.: Haptic Science Learning System for Students with Visual Impairments: A Preliminary Study. In: Proceedings of the Xth International Conference on Human-Computer Interaction (HCI 2009), USA (2009)
15. Akahane, K., Yamaguchi, T., Isshiki, M., Sato, M.: Promoting Haptic Interaction Technology - SPIDAR-G & SPIDAR-mouse. In: EuroHaptics 2010, Amsterdam (2010)

Part II

Gaze and Gesture-Based Interaction

Face Recognition Using Local Graph Structure (LGS)

Eimad E.A. Abusham and Housam K. Bashir

Faculty of Information Science and Technology, Multimedia University
Melaka, 75450, Malaysia
eimad.eldin@mmu.edu.my, me.the.fren@gmail.com.

Abstract. In this paper, a novel algorithm for face recognition based on Local Graph Structure (LGS) has been proposed. The features of local graph structures are extracted from the texture in a local graph neighborhood then it's forwarded to the classifier for recognition. The idea of LGS comes from dominating set points for a graph of the image. The experiments results on ORL face database images demonstrated the effectiveness of the proposed method. The advantages of LGS, very simple, fast and can be easily applied in many fields, such as biometrics, pattern recognition, and robotics as preprocessing.

Keywords: Algorithm, Feature evaluation and selection, Pattern Recognition, Pattern Recognition.

1 Introduction

The question of whether face recognition is done holistically or using the local feature analysis has also been researched in the literature. Studies done by Bruce [1] suggested the possibility of global descriptions or holistic representations as a precursor or pre-process for finer feature recognition. Various approaches in face recognition have been proposed in the literature; these can be classified into three categories, namely feature-based, holistic (global), and hybrid methods. While feature-based approaches compare the salient facial features or components detected from the face, holistic approaches make use of the information derived from the whole face pattern. By combining both local and global features, the hybrid methods attempt to produce a more complete representation of facial images.

The main idea behind this feature-based technique is to discriminate among the different faces based on the measurement of structural attributes of the face. The method extracts and computes a set of geometrical features from faces such as the eyes, eyebrows, nose, and mouth and feeds them into a structural classifier. One of the earliest methods, included in this category, was proposed by Kelly [2]. In this method, Kelly used the width of the head, the distances between the eyes and from the eyes to the mouth, etc. Kanade [3] applied the distances and angles between eye corners, mouth extrema, nostrils, and chin top in his work on Computer Recognition of Human Faces. Cox, Ghosn and Yianios [4] introduced a mixture-distance technique, which used manually extracted distances, with each 30 manually extracted distances that represented face. Without finding the exact locations of the facial features, Hidden Markov Model (HMM) has been successfully applied to face recognition. The HMM

based methods use strips of pixels that cover the forehead, eyes, nose, mouth, and chin [5-6]. [5]Nefian yielded a better performance than Samaria using the KL projection coefficients instead of the strips of raw pixels.

One of the most successful systems in the graph matching system reported by Wiskott et al. [7] employs the Gabor wavelet features at facial landmarks. A flexible face structure or elastic bunch graph match (based on the dynamic link architecture or DLA), presented in Buhmann, Lades & van der Malsburg [8] and Lades et al. [9], is combined with the local Gabor features description onto one system. Elastic Bunch Graph Matching, a topology graph was constructed for each face first, with each node attaching one or several Gabor jets. Each component of a jet is a filter response of a specific Gabor wavelet, extracted at a pre-defined critical feature point.

In the Elastic Bunch Graph Matching, a topology graph is first constructed for each face, with each node attaching one or several Gabor jets. Each component of a jet is a filter response of a specific Gabor wavelet, extracted at a pre-defined critical feature point. These locally estimated Gabor features are known as robust against illumination change, distortion and scaling [10], and this is the first key factor in the EBGM method. Another key point of this method lies in the graph matching, of which the first step is similar to that in Lee [10], i.e., both the local and global similarities are considered. The second step, where a deformable matching mechanism is employed, i.e., each node of the template graph is allowed to vary its scale and position according to the appearance variations on a specific face. To investigate the discriminative power of the Gabor features (nodes), a systematic way of selecting the nodes from a dense set is presented in Krüger [11] and Krüger et al.[12] . In their work, more nodes were found to yield better results, because more information was used. Nevertheless, this effect saturates if the nodes are too close and the corresponding Gabor coefficients become highly correlated due to the overlap between the kernels. On the other hand, the computational effort linearly increases with the number of nodes. The optimal number of nodes becomes a compromise between recognition performance and speed. The EBGM method has been proven to be very effective in face recognition and was one of the top performers in the FERET evaluation tests. Learning discriminative facial locations and obtaining optimal local feature extractor parameters are formulated as a feature subset selection problem. In feature selection, the aim is to select a subset from a given set such that the classification accuracy of the selected subset is maximized.

Baldi and Hornik [13] generates an optimal linear encoding using optimization methods in layered linear feed-forward neural networks to neutrally spanned subspace, average distortion is minimized by using the principal components of the examples. pairwise relationships between pixels of the image is computed by PCA and important information which contained in high-order relationship among pixels are discarded. Therefore, it reasonable to look for a method that considers this high-order statistic. One of these methods is Independent component analysis (ICA) uses sigmoidal neurons to derive the principle of optimal information transfer, Movellan & Sejnowski[14]. In 2007, Zou, Ji, & Nagy [15] have conducted comparative studies on local matching approaches. The general idea of local matching methods is to first locate several facial features and then classify the faces by comparing and combining the corresponding local statistics. Ruiz-del-Solar, Verschae, Correa, [16] have studied and analyzed four face recognition methods that are based on different representations

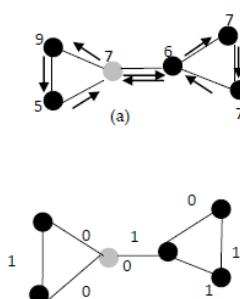
of the image: (1) LBP Histograms, (2) Gabor Jet Descriptors (GJD), (3) SIFT Descriptors, and (4) ERCF (Extremely Randomized Clustering Forest) of SIFT Descriptors. These representations are used in different ways by the analyzed face recognition methods: (1) LBP Histograms are directly used as feature vectors together with distance metrics for comparing these histograms, (2) GJD are used together with Borda count, (3) SIFT Descriptors are used together with local and global matching methods, and (4) ERCF are used together with linear classifiers.

The studies mentioned above made use of some forms of local information in their systems. These local regions are either rigidly chosen or placed over facial landmarks. To some extent, the selection of these regions is based on an intuitive knowledge of facial structure and the importance of facial features. However, these features may not always be optimal in building a face representation for the task of recognition. Instead, it may be better to learn an optimal selection of the features directly from the data rather than using a manual selection.

2 Local Graph Structure (LGS)

The idea of Local Graph Structure (LGS) comes from a dominating set for a graph $G = (V, E)$ is a subset D of V such that every vertex not in D is joined to at least one member of D by some edge. The domination number $\gamma(G)$ is the number of vertices in a smallest dominating set for G .

LGS works with the six neighbors of a pixel, by choosing the target pixel C as a threshold, then we start by moving anti clockwise at the left region of the target pixel C, If a neighbor pixel has a higher gray value than the target pixel (or the same gray value) then assign a binary value equal to 1 on the edge connecting the two vertices, else we assign a value equal to 0. After finish on the left region of graph we stop at the target pixel C and then we move in a horizontal way (clockwise) to the right region of the graph and we apply the same process till we get back to the target pixel C.



Binary: 01010110 - Decimal: 86

Fig. 1. Local graph structure (a. Direction, b. Binary)

To produce the LGS for pixel (x_d, y_d) a binomial weight 2^p is assigned to each sign $s(g_d - g_n)$. These binomial weights are summed:

$$LGS(x_d, y_d) = \sum_{K=0}^7 s(g_d - g_n) 2^K \quad (1)$$

$$\text{where } s(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Where $p = 7$. $p = 7, 6, \dots, 0$.

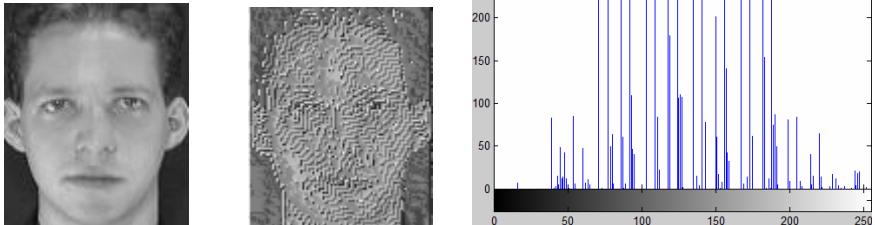


Fig. 2. LGS operator

3 Experiments and Results

Local Graph Structure (LGS) have proved to be useful because it contains information about the distribution of the local micropatterns, such as edges, spots and flat areas, over the whole image. A decimal representation is obtained by taking the binary sequence as a binary number between 0 and 255. For dominant pixel, not only accounts for its relative relationship with its neighbours but also the relationship between the pixels that form the local graph of the target pixel C (dominant pixel), while discarding the information of amplitude, and this makes the resulting LGS values very insensitive to illumination intensities. The 8-bit binary series with binomial weights consequently result in 256 different patterns in total for the pixel representation.

In the initial work of face processing using LGS, can be seen in Fig 3, is an example of new generated image from original image using LGS, a histogram of the LGS representing the distribution of 256 patterns across the face image. The

advantage of LGS; Firstly it is a local measure; some of the regions contain more useful information than others when face image been divided into regions, it can be expected to assist in terms of distinguishing between people. Secondly it is a relative measure, and is invariant to any monotonic transformation such as shifting, scaling, or logarithm of the pixel-values. Therefore it can be invariant to a certain range of illumination changes.

We test our method using the public ORL face database [17]. The database consists of 400 faces; Ten different images of each of 40 distinct subjects. The faces were captured with the subjects in a straight, frontal position against a dark identical background, and with acceptance for some sloping and regular change of up to 20 degrees. Image variations of five individuals in the database are illustrated in Fig3.



Fig. 3. Example of an original image of ORL face database

To assess the performance of LGS on face recognition, 40 subjects have been taken for our experiments, 8 images for training and the remaining 2 images for testing; LGS applied to find the histograms for the entire training and testing sets. The correlation is used to computes the correlation coefficient of histogram between two images recognition. For e.g. A and B are two different histogram of images, A and B are vectors of the same size. The correlation coefficient is computed as follows:

$$result = \frac{\sum \sum (A_{mn} - \bar{A})(B_{mn} - \bar{B})}{\sqrt{\left[\sum \sum (A_{mn} - \bar{A}^2) \right] \left[\sum \sum B_{mn} - \bar{B}^2 \right]}} \quad (2)$$

Table 1 illustrates the sample of result obtained by the proposed method.

Table 1. Similarity Rate

Subjects	Testing Image	Index of output	Similarity (with Training) ALG
Subject 1	1	1	99.22%
	2	5	98.01%
Subject 2	3	10	99.67%
	4	12	99.65%
Subject 3	5	17	99.63%
	6	17	99.53%
Subject 4	7	30	99.59%
	8	25	99.68%
Subject 5	9	35	99.66%
	10	40	99.67%
.....
Subject 40	79	315	99.70%
	80	319	99.53%

The overall detection rate of LBGS is shown in Table 2.

Table 2. Overall Recognition RATE

LGS	Recognition Rate
Overall	93.75%
Max Recognition	99.87%
Min Recognition	98.01%

4 Conclusion

The features of local binary graph structure are derived from a general definition of texture in a local graph neighborhood. The advantages of LGS over other local methods it's invariant to illumination changes, computational efficiency, and fast so that it can be easily applied in real-time system. LGS assigns weight for target pixels (dominant) by considering not only the direct relationship of target pixels to its neighbours but also the relationship between the pixels that form the local graph of the target pixel; this feature is unique to LGS and lead to improve the image appearance and subsequently the recognition performance. This is especially applicable for faces from which the photos are taken under different lighting conditions. Important regions for feature extraction are those with the eyes and the mouth as you can see in figure. ALG can easily be combined with other methods and can easily kernelized by using different kernels functions.

Acknowledgment

The authors thank the Olivetti Research Laboratory (United Kingdom) for using their Olivetti database (formally known ORL database).

References

- [1] Bruce, V., et al.: Human face perception and identification. NATO ASI series. Series F: Computer and System Sciences, pp. 51–72 (1998)
- [2] Kelly, M.: Visual identification of people by computer. Dept. of Computer Science, Stanford University (1970)
- [3] Kanade, T.: Computer recognition of human faces (1977)
- [4] Cox, I., et al.: Feature-based face recognition using mixture-distance, Computer Vision and Pattern Recognition. IEEE Press, Piscataway (1996)
- [5] Nefian, A., Iii, M.: A hidden Markov model-based approach for face detection and recognition (1998)
- [6] Samaria, F., Young, S.: HMM-based architecture for face identification. Image and Vision Computing 12, 537–543 (1994)
- [7] Wiskott, L., et al.: Face recognition by elastic bunch graph matching. IEEE Transactions on Pattern Analysis and Machine Intelligence 19, 775–779 (2002)
- [8] Buhmann, J., et al.: Size and distortion invariant object recognition by hierarchical graph matching, pp. 411–416 (2002)
- [9] Lades, M., et al.: Distortion invariant object recognition in the dynamic link architecture. IEEE Transactions on Computers 42, 300–311 (2002)
- [10] Lee, T.: Image representation using 2D Gabor wavelets. IEEE Transactions on Pattern Analysis and Machine Intelligence 18, 959–971 (2002)
- [11] Krüger, N., et al.: Determination of face position and pose with a learned representation based on labelled graphs*. 1. Image and Vision Computing 15, 665–673 (1997)
- [12] Krüger, N., et al.: Autonomous learning of object representations utilizing self-controlled movements. In: Proc. of Neural Networks in Applications, NN, vol. 98, pp. 25–29 (1998)
- [13] Baldi, P., Hornik, K.: Neural networks and principal component analysis: Learning from examples without local minima. Neural networks 2, 53–58 (1989)
- [14] Bartlett, M., et al.: Face recognition by independent component analysis. IEEE Transactions on Neural Networks 13, 1450–1464 (2002)
- [15] Zou, J., et al.: A comparative study of local matching approach for face recognition. IEEE Transactions on Image Processing 16, 2617–2628 (2007)
- [16] Javier, R., et al.: Recognition of faces in unconstrained environments: a comparative study. EURASIP Journal on Advances in Signal Processing 2009 (2009)
- [17] Samaria, F., Harter, A.: Parameterisation of a stochastic model for human face identification, pp. 138–142 (1994)

Eye-gaze Detection by Image Analysis under Natural Light

Kiyohiko Abe¹, Shoichi Ohi², and Minoru Ohyama²

¹ College of Engineering, Kanto Gakuin University, 1-50-1 Mutsuura-higashi,
Kanazawa-ku, Yokohama-shi, Kanagawa 236-8501, Japan

² School of Information Environment, Tokyo Denki University, 2-1200 Muzaigakuendai,
Inzai-shi, Chiba 270-1382, Japan
abe@kanto-gakuin.ac.jp

Abstract. We have developed an eye-gaze input system for people with severe physical disabilities, such as amyotrophic lateral sclerosis (ALS). The system utilizes a personal computer and a home video camera to detect eye-gaze under natural light. Our practical eye-gaze input system is capable of classifying the horizontal eye-gaze of users with a high degree of accuracy. However, it can only detect three directions of vertical eye-gaze. If the detection resolution in the vertical direction is increased, more indicators will be displayed on the screen. To increase the resolution of vertical eye-gaze detection, we apply a limbus tracking method, which is also the conventional method used for horizontal eye-gaze detection. In this paper, we present a new eye-gaze detection method by image analysis using the limbus tracking method. We also report the experimental results of our new method.

Keywords: Eye-gaze detection, Image analysis, Natural light, Limbus tracking method, Welfare device.

1 Introduction

Recently, eye-gaze input systems were reported as novel human-machine interfaces [1], [2], [3], [4], [5]. Users can use these systems to input characters or commands to personal computers. These systems require only the eye movement of the user as an input. In other words, the operation of these systems involves the detection of the eye-gaze of the users. As in the case of our study, these systems have been used to develop communication aid systems for people with severe physical handicaps such as amyotrophic lateral sclerosis (ALS).

We have developed an eye-gaze input system [4], [5]. The system utilizes a personal computer and a home video camera to detect eye-gaze under natural light. Eye-gaze input systems usually employ a non-contact-type eye-gaze detection method. Natural light (as well as artificial light sources such as fluorescent lamps) can be used as a light source for eye-gaze detection. The eye-gaze input systems that operate under natural light often have low accuracy. Therefore, they are capable of classifying only a few indicators [2]. To resolve this problem, a system using multi-cameras is proposed [3].

We have developed a new eye-gaze input system, which employs multi-indicators (27 indicators in 3 rows, and 9 columns) [4]. This system comprises a personal computer and a home video camera. In other words, the system is not only inexpensive but also user friendly; therefore, it is suitable for personal use such as in welfare device applications. In addition, we developed an application system for our eye-gaze input system that supports personal computers (Microsoft Windows XP or Vista), English and Japanese text input, web browsing, etc. [5].

The practical eye-gaze input system described above is capable of classifying the horizontal eye-gaze of users with a high degree of accuracy [4]. However, it can detect only three directions of vertical eye-gaze. If the detection resolution in the vertical direction is increased, more indicators will be displayed on the screen. This factor is an advantage when designing a more user-friendly interface. The conventional method developed by us for vertical eye-gaze detection is based on a similar method that uses light intensity distributions (the results of a one-dimensional projection) of an eye image. Therefore, if the resolution of the vertical eye-gaze detection increases, many eye images will be required as reference data.

To improve this point, we apply a limbus tracking method for vertical eye-gaze detection, which is also the conventional method used for horizontal eye-gaze detection. In other words, we developed a detection method for obtaining the coordinate data of the user's gazing point. This method arranges its detection area on the open-eye area of the eye image. The eye-gaze of the user is estimated by the integral values of the light intensity on the detection area.

2 Eye-gaze Detection by Image Analysis

The aim of the eye-gaze system is to detect the eye-gaze of the user. Several eye-gaze detection methods have been studied. To detect eye-gaze, these methods analyze eye images captured by a video camera [1], [2], [3], [4], [5]. This method of tracking the iris by image analysis is the most popular detection method that is used under natural light [2], [3]. However, it is difficult to distinguish the iris from the sclera by image analysis, because of the smooth transition of the luminance on between the iris and the sclera. In some users, the iris is hidden by their eyelids, which makes it difficult to estimate the location of the iris by elliptical approximation.

We propose a new eye-gaze detection method that involves image analysis using the limbus tracking method [4], [5]. This method does not estimate the edge of the iris. Here, we describe the eye-gaze input detection method in detail. The location and the size of detection area is fixed, as described below.

An overview of the horizontal eye-gaze detection method developed by us is shown in Fig. 1 (a). The difference in the reflectance between the iris and the sclera is used to detect the horizontal eye-gaze. In other words, the horizontal eye-gaze is estimated by using the difference between the integral values of the light intensity on areas A and B, as shown in Fig. 1 (a). We define this differential value as the horizontal eye-gaze value.

An overview of the proposed vertical eye-gaze detection method is shown in Fig. 1 (b). Vertical eye-gaze is also based on the limbus tracking method. We estimate the vertical eye-gaze by using the integral value of the light intensity on area C, as shown in Fig. 1 (b). We define this integral value as the vertical eye-gaze value. If the eye-gaze input system is calibrated using the relations between these eye-gaze values and the angle of sight, we can estimate the horizontal and vertical eye-gaze of the user. The details of the calibration method are described in chapter 4.

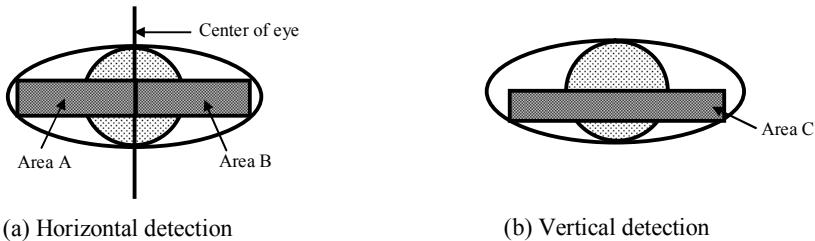


Fig. 1. Overview of eye-gaze detection

3 Automatic Arrangement of Detection Area

We are developing an eye-gaze input system, as described above. By using this system, users can select icons on the computer screen or control the mouse cursor by eye-gaze. In general, if the user's eye-gaze moves, the shape of the eyelids (upper or lower) change. We have observed that the error of detection increases with a change in the shapes of the eyelids because the shape of the detection area also changes. To resolve this problem, we developed a new method for estimating the detection area. This method extracts the exposed eye area from an eye image by analyzing the part of the image where the exposed eye area is not hidden by the eyelids. The detection area can be used to detect both the horizontal and the vertical eye-gaze.

3.1 Extraction of Exposed Eye Area

We can estimate the exposed eye area by merging the open-eye areas. The merging method uses a bitwise AND operation. The open-eye areas are extracted from eye images when the user directs his or her gaze at indicators 1 to 9, shown in Fig. 2. These open-eye areas are extracted by binarization using the color information of the skin [4]. To reduce the calculation cost without losing accuracy, we conducted pre-experiments that use some groups of indicators. Here, we used the open-eye areas extracted by the four eye images when the user directs his or her gaze at indicator 2 (up), 4 (left), 6 (right), and 8 (down). We use the exposed eye area estimated from these open-eye areas for extracting the detection area for the horizontal and vertical eye-gaze. An overview of the exposed eye area extraction is shown in Fig. 3.

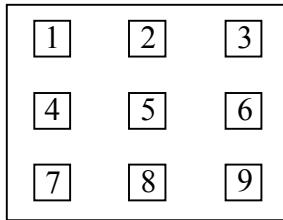


Fig. 2. Gaze indicators

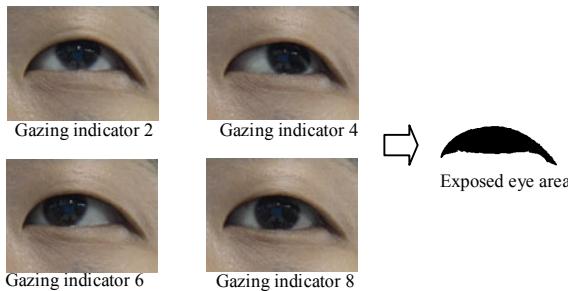


Fig. 3. Extraction of exposed eye area

3.2 Estimation of Detection Area

To utilize the image analysis based on the limbus tracking method, we have to focus the detection area on the pixels whose light intensity changes. Therefore, we estimate the detection area of the vertical and horizontal eye-gaze by using the eye images when a user looks at indicator 2 (up) and 8 (down), or indicator 4 (left) and 6 (right), respectively. First, we create an image that consists of the difference in the images captured when a user looks at indicator 2 and 8. We estimate the detection area of the vertical eye-gaze from the created image. In practical terms, the detection area for the vertical eye-gaze is arranged on the pixels with positive differences values. This process is executed inside the exposed eye area, as shown in Fig. 3.

Next, we estimate the centerline of the detection area using the horizontal eye-gaze detection method, shown in Fig. 1 (a). In practical terms, this centerline is located where the horizontal eye-gaze value is at a maximum. We estimate the horizontal eye-gaze from the eye images when a user looks at indicators 4 and 6. This process is executed inside the detection area of the vertical eye-gaze. Fig. 4 shows a sample of the detection area estimated by our proposed method.

We estimate the vertical eye-gaze by using the integral value of the light intensity on the total detection area. We estimate the horizontal eye-gaze from the difference between the integral values of the light intensity on the two detection areas split by the centerline.

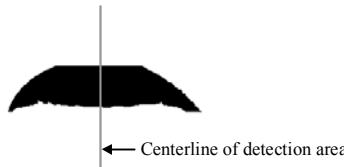


Fig. 4. Detection area estimated by our proposed method

4 Calibration

4.1 Calibration for Eye-gaze Detection

The characteristics of eye-gaze detection of our proposed method are different for different users. Therefore, the eye-gaze input system needs to be calibrated for eye-gaze detection. For typical calibration, the user looks at the indicators arranged at regular intervals on the computer screen. The system is calibrated by using these results. The characteristics of the vertical and horizontal eye-gaze are shown in Fig. 5 as a scatter plot. In Fig. 5, the abscissa axis and longitudinal axis indicate the horizontal and vertical eye-gaze value respectively. In addition, the indicator numbers (Ind. 1 to 9) are displayed near the plot points. These correspond to the indicator numbers shown in Fig. 2.

From Fig. 5, it is evident that if the eye-gaze of the user moves only in a horizontal direction, the vertical eye-gaze values do change as well. For example, indicators 1 to 3 are arranged on the same horizontal line, but the estimated vertical eye-gaze values have different values. Similarly, it is also evident that if the user's eye-gaze moves in a vertical direction only, the horizontal eye-gaze values change as well.

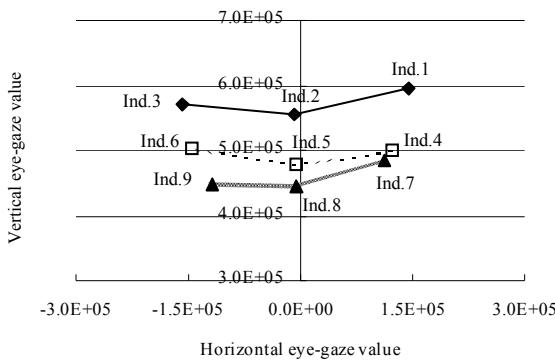


Fig. 5. Characteristics of vertical and horizontal eye-gaze (Subject A)

4.2 Calibration Method by Interpolation between Indicators

There is a dependency relationship between the characteristics of the horizontal and the vertical eye-gaze, as described in section 4.1. Considering this point, we calibrate

our eye-gaze input system by using the four indicator groups. The indicators are separated into four groups, for example, “indicator 1, 2, 4, 5 (upper left group)”, “indicator 2, 3, 5, 6 (upper right group)”, “indicator 4, 5, 7, 8 (lower left group)”, and “indicator 5, 6, 8, 9 (lower-right group)”. The eye-gaze input system is calibrated by using each indicator group. An overview of the calibration method using indicators 1, 2, 4, 5 is shown in Fig. 6.

First, we note the characteristics of the horizontal eye-gaze. The characteristics from indicators 1 and 2 and indicators 4 and 5 are defined as Line 1 and Line m, respectively. From Fig. 6, it is evident that the gradient of Line 1 is greater than Line m. If the user’s eye-gaze moves down, the horizontal eye-gaze value increases. The calibration method for horizontal eye-gaze detection calculates the change in the parameters of these calibration functions to calculate their gradients. We assume that the changes in the parameters are proportional to the horizontal eye-gaze movement. We estimate the calibration function for the vertical eye-gaze detection using a similar method. If we calculate the calibration functions by using the horizontal and vertical eye-gaze values, ($V_{h1,2,4,5}$, $V_{v1,2,4,5}$), while the user looks at the indicators (Ind. 1, 2, 4, 5), we can estimate where the user is looking.

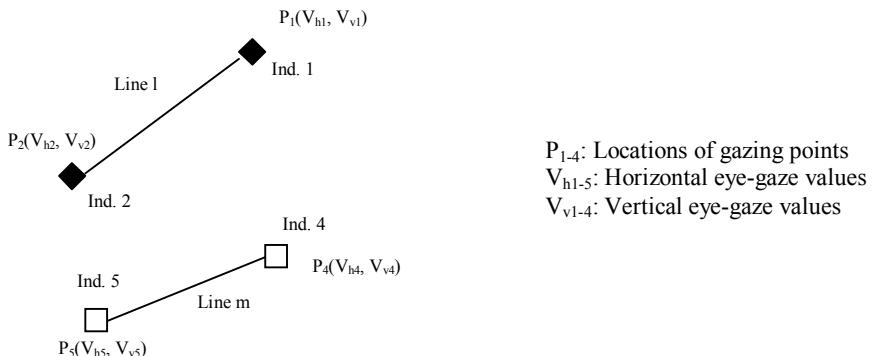


Fig. 6. Calibration method by interpolation between indicators

5 Evaluation Experiments for the Proposed Method

Evaluation experiments were conducted with five subjects. In the evaluation experiments, we detected the horizontal and vertical eye-gaze of each subjects, and then estimated the gaze-detection errors. The evaluation system was calibrated by using the method described in section 4.

5.1 Overview of the Experiment System

The evaluation experiment system comprises a personal computer, a home video camera, and an IEEE1394 interface for capturing the images from the video camera. The experiments were conducted under general room lighting (fluorescent light). The hardware configuration of the experimental setup is shown in Fig. 7. The video camera records images of the user’s eye from a distant location (the distance between

the user and camera is approximately 70 cm) and then this image is enlarged. The user's head movements induce a large error in the detections. We compensated for the head movement by tracing the location of an inner corner within the eye [4], [5].

Subjects must calibrate the system before using it, as described above. While the calibration is being performed, subjects look at each of the nine indicators shown in Fig. 8 (a). If the calibration terminates successfully, the subject then looks at the indicators shown in Fig. 8 (b). We estimate the gaze-detection errors as described earlier. The indicators are square and 1.5(deg.) in size. The maximum breadths of displayed indicators are 18(deg.) (vertical) and 24(deg.) (horizontal). These sizes are shown as angle of sight.

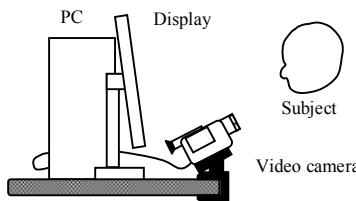


Fig. 7. Hardware configuration of experiment system

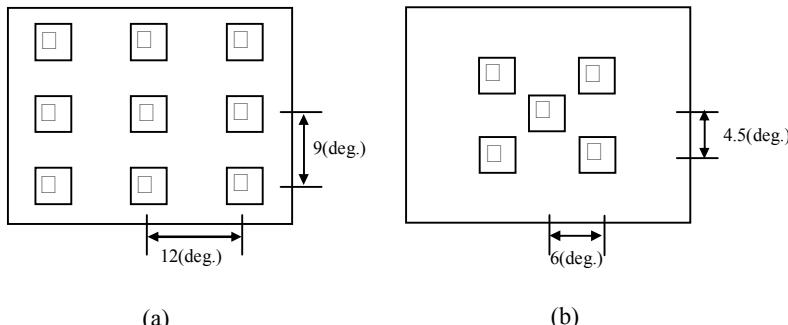


Fig. 8. Indicators for evaluation experiments

5.2 The Detection Errors of Gaze Points

Tables 1 and 2 show the gaze-detection errors of the five subjects and the results of the horizontal and vertical eye-gaze detection. It is evident that the horizontal detection errors are smaller than the vertical detection errors. We confirmed the trend for all the subjects. In practical terms, the results of subjects A, B, D, and E indicate that the errors of vertical detection are over 2(deg.). These errors are larger than the errors of horizontal detection. We confirmed that the dynamic range of the vertical eye-gaze value is narrower than the horizontal value. In the results, the negative effect of the noise on the eye image is increased. The change in illumination on the experimental conditions cause this noise.

Table 1. Detection errors of gazing points(deg.) (horizontal)

	Subject A	Subject B	Subject C	Subject D	Subject E
Ind. 1	0.29	0.11	0.72	0.33	0.53
Ind. 2	0.82	0.53	0.52	0.62	0.88
Ind. 3	0.09	0.67	0.81	1.33	0.35
Ind. 4	0.19	0.76	1.23	0.37	0.32
Ind. 5	0.50	0.01	0.33	0.82	0.78
Average	0.38	0.42	0.72	0.69	0.57
Standard variation	0.29	0.34	0.34	0.41	0.25

Table 2. Detection errors of gazing points(deg.) (vertical)

	Subject A	Subject B	Subject C	Subject D	Subject E
Ind. 1	2.53	0.20	0.62	1.53	2.06
Ind. 2	0.38	2.79	0.60	0.63	1.06
Ind. 3	1.76	1.48	0.25	1.37	0.61
Ind. 4	0.03	0.49	0.65	0.04	2.02
Ind. 5	0.07	0.16	1.31	2.93	1.78
Average	0.95	1.02	0.69	1.30	1.51
Standard variation	1.13	1.12	0.39	1.09	0.64

We can summarize the obtained results as follows: From Tables 1 and 2, it is evident that the average errors of vertical and horizontal eye-gaze detection are 1.09(deg.) and 0.56(deg.), respectively. These results indicate that our proposed method can detect vertical and horizontal eye-gaze to a high degree of accuracy, and performs as well as the detection method that uses infrared light [6]. Our conventional method is capable of classifying the horizontal eye-gaze of users with a high degree of accuracy. However, it can detect only three directions in the vertical eye-gaze. Therefore, the conventional eye-gaze input system is capable of classifying 27 indicators (in 3 rows and 9 columns). If our proposed method is used in an eye-gaze input system, the system can classify approximately 45 indicators because it can detect five indicators in the vertical eye-gaze.

6 Conclusions

We present a new eye-gaze input system using image analysis based on the limbus tracking method. This system uses a personal computer and a home video camera to detect eye-gaze under natural light. The eye-gaze detection method for our proposed system is simple. In other words, this method does not need special devices, such as infrared light. Therefore, the size of this system is small and it is highly versatile.

Our conventional eye-gaze detection method can detect the horizontal eye-gaze with a high degree of accuracy. However, it can only detect three directions in the vertical eye-gaze. Our proposed method can detect both the horizontal and the vertical eye-gaze with a high degree of accuracy. Therefore, by using our proposed method, we can estimate where the user is looking in a two-dimensional plane.

The evaluation experiments for our proposed method were conducted with five subjects. The results for the five subjects show that the average detection errors of vertical and horizontal gaze points are approximately 0.56(deg.) and 1.09(deg.), respectively. These results indicate that the eye-gaze input system using the method developed by us is capable of classifying approximately 45 indicators. In other words, the system can classify nearly double the number of indicators.

In future studies, we will develop a new detection method for vertical eye-gaze, and enhance its detection accuracy. Furthermore, we will develop a more user-friendly eye-gaze input system by using the newly developed method for eye-gaze detection.

References

1. Hutchinson, T.E., White Jr., K.P., Martin, W.N., Reichert, K.C., Frey, L.A.: Human-computer Interaction using Eye-gaze Input. *IEEE Trans. Systems, Man, and Cybernetics* 19(7), 1527–1534 (1989)
2. Corno, F., Farinetti, L., Signorile, I.: A Cost-effective Solution for Eye-gaze Assistive Technology. In: Proc. IEEE International Conf. on Multimedia and Expo, vol. 2, pp. 433–436 (2002)
3. Wang, J.G., Sung, E.: Study on Eye-gaze Estimation. *IEEE Trans. on Systems, Man and Cybernetics* 32(3), 332–350 (2002)
4. Abe, K., Ohyama, M., Ohi, S.: Eye-gaze Input System with Multi-Indicators Based on Image Analysis under Natural Light (in Japanese). *J. The Institute of Image Information and Television Engineers* 58(11), 1656–1664 (2004)
5. Abe, K., Ohi, S., Ohyama, M.: An Eye-gaze Input System Using Information on Eye Movement History. In: Proc. on 12th International Conference on Human-Computer Interaction, HCI International 2007, Beijing, vol. 6, pp. 721–729 (2007)
6. Ramdane-Cherif, Z., Nait-Ali, A.: An Adaptive Algorithm for Eye-gaze-Tracking-Device Calibration. *IEEE Trans. on Instrumentation and Measurement* 57(4), 716–723 (2008)

Multi-user Pointing and Gesture Interaction for Large Screen Using Infrared Emitters and Accelerometers

Leonardo Angelini, Maurizio Caon, Stefano Carrino,
Omar Abou Khaled, and Elena Mugellini

College of Engineering and Architecture of Fribourg, Switzerland
{Leonardo.Angelini,Maurizio.Caon,Stefano.Carrino,
Omar.AbouKhaled,Elena.Mugellini}@Hfr.ch

Abstract. This paper presents PlusControl, a novel multi-user interaction system for cooperative work with large screen. This system is designed for a use with economic deictic and control gestures in air and it allows free mobility in the environment to the users. PlusControl consists in light worn devices with infrared emitters and Bluetooth accelerometers. In this paper the architecture of the system is presented. A prototype has been developed in order to test and evaluate the system performances. Results show that PlusControl is a valuable tool in cooperative scenarios.

Keywords: human computer interaction, large screen, gesture recognition, visual tracking, computer supported cooperative work, economic gestures.

1 Introduction

In 1980, A. Bolt introduced his Put-That-There system to the scientific community [1]. He started an important revolution in the Human-Computer Interaction area; in fact, he stated that the machine should understand the human language and not the contrary, making interaction between human and machine more natural. The natural way of interaction between human beings is based on speech and gestures. Therefore, in order to make human computer interaction more natural, a system should get closer to these forms of multimodal communication [2]. Unfortunately the voice commands are not convenient in any situation. For example, in a conference scenario there could be some troubles with the speech of the people making a presentation. In fact some words could be incorrectly understood by the computer as a voice command. Adding a gesture as trigger for the speech command could be a solution but such a way of interaction introduces a break in the presentation. Our aim, instead, is to make the interaction and the presentation two seamless and synergic experiences. Therefore a system that involves exclusively gesture recognition should be preferable.

The support for multi-user interaction can be really important in areas that involve cooperative work such as for instance educational application. An example of a cooperative work application can be found in [3], in which new devices (e.g. IntelliPen, a laser tracked pen for direct manipulation of objects on the projection wall) and interaction modalities (e.g. gestures) have simplified and accelerated the engineering and design development phases. Moreover, single-display groupware

systems have been discovered as an important tool to make students more motivated, more involved in the class and more socially interactive [4]. At present, those systems utilize several mice for the interaction but a free air system should be more comfortable. Ying Tin and Randall Davis stated that interfaces closely conformable to the way people naturally interact would have the potential to lower the users' cognitive load, allowing them to concentrate on the decision-making task. This is a feature that should be precious in an educational context where the users are supposed to interact focused on contents. The same concept has been applied in the creation of the USAR (urban search and rescue) system to help people in organizing and coordinating rescue tasks in a high pressure context, like aiding victims in the aftermath of a disaster [5].

The system presented in this paper is based on the recognition of pointing and control gestures performed in air granting simultaneous interaction of multiple users on a large screen. PlusControl is designed for natural and economic interaction, facilitating its learnability.

Reaching large screen with small movements is not easy without losing precision, but the developed system exploits the speed of hand movements to grant higher precision and smaller displacement with slow gestures, and larger displacements and lower precision of the cursor with quick gestures. Moreover the interaction has been designed to perform gestures which require small movements in order to make a less tiring experience [6].

This system combines the advantages of an easy-to-use pointing system with a natural gesture language for commands.

In order to provide a clear research context for PlusControl, the related works are analysed in Section 2. In Section 3, an overview of the PlusControl architecture is presented. In Sections 4 and 5, the testing methodology and the obtained results are shown. Section 6 concludes the paper and presents future works.

2 Related Works

Interaction with large screens is a demanding area for research because of the wide diffusion of very large displays and projectors. In fact realizing an interface for these applications leads to many serious technical problems. For example, the large size of these screens can cause users to lose track of the mouse pointer and it can make really difficult reaching distant contents [7]. To cope with these problems, many systems, based on different technologies, have been developed. Some systems have touch interfaces (e.g. [8]) but they require the users to remain near to the displays. Therefore, to grant free mobility to the users in the environment while maintaining the possibility to continue the interaction with the system, many researchers adopted different solutions for pointing. Most of them have chosen laser pointers as cursors [9], [10], [11]. This technique is based on computer vision: the laser dot on the screen is captured with a camera and the image is processed to obtain the cursor information. Using only the pointing system, it is possible to control an interface but it is necessary to introduce complex solutions to realize even the simpler commands like clicking. For example, in [12], a LED source is tracked by a camera for pointing; the clicking action is obtained by the double crossing interaction. In the article [13] the authors put a button on the laser pointer to click. Jiang et al. made a pointer handheld device

combining a camera with the Logitech Cordless Presenter [14]. Clicking is an essential feature for most applications, but a pointing system could be augmented by gestures to execute more actions with free mobility, granting a seamless experience. In the gesture recognition domain, related works are innumerable. There are many technologies that allow movements capturing; among them, accelerometers are really quoted by researchers since they are already present in several devices, have a low price and achieve excellent performances [15], [16], [17], [18], [19].

Many systems allow pointing and clicking granting free mobility; others implement gesture recognition for a natural interaction. PlusControl is a novel system that combines these two features adding the possibility of multi-user interaction for cooperative work.

3 PlusControl Architecture

The PlusControl system is designed to support concurrent multi-user interaction on a large screen. In order to allow to the users a seamless and natural cooperation, the system showed in Fig. 1 has been developed. The interaction system is composed of two accelerometers and an infrared emitter for each person, which can be worn on a finger or held in the hand. The infrared emitter is tracked with two IR cams. The two cameras are horizontally parallel; they are positioned turned towards the users and under the large screen. The small control movements allow using only a part of the range of vision of the cameras. Moreover, two parallel cameras grant a wider area of interaction, thus a multi-user control is possible. Accelerometers are used to combine the cursor movement with both simple gestures (e.g. press click, release click, simple click) and more complex gestures (e.g. user recognition, previous slide, next slide).

The algorithm that calculates the displacement of the cursor using data from the IR emitters tracking has been improved with a distance factor correction based on stereo vision triangulation. In fact, using two IR cameras it is possible to calculate the approximated distance of the user by measuring the horizontal disparity of the two blobs tracked from each of the two cameras. The correction factor makes the interaction not dependent from the distance of the user from the cameras. This approach grants small movements of the hand also at large distances, improving the usability.

The main difficulty in a multi-user environment is to associate the correct tracked blob to each user. The user recognition is done with a particular gesture that allows matching the accelerometer data with the tracked blob data. A quick up-down movement has been chosen as gesture for the user recognition (assigning the accelerometers data to the tracked blob) when his emitter enters the range of vision of the camera.

Using the distance correction system, the blobs to be assigned to each user are two. As a consequence even if one of the two cameras loses its blob, the system is still able to move the cursor using the other blob but the distance of the user will not be calculated. Until one of the two blobs is tracked, the system is able to re-associate the second blob when it reaches again the field of vision of the cameras. If both the blobs are lost, the user should do again the gesture to be recognized from the system (i.e. the quick up-down movement).

In addition working with infrared light grants better robustness against light changes.

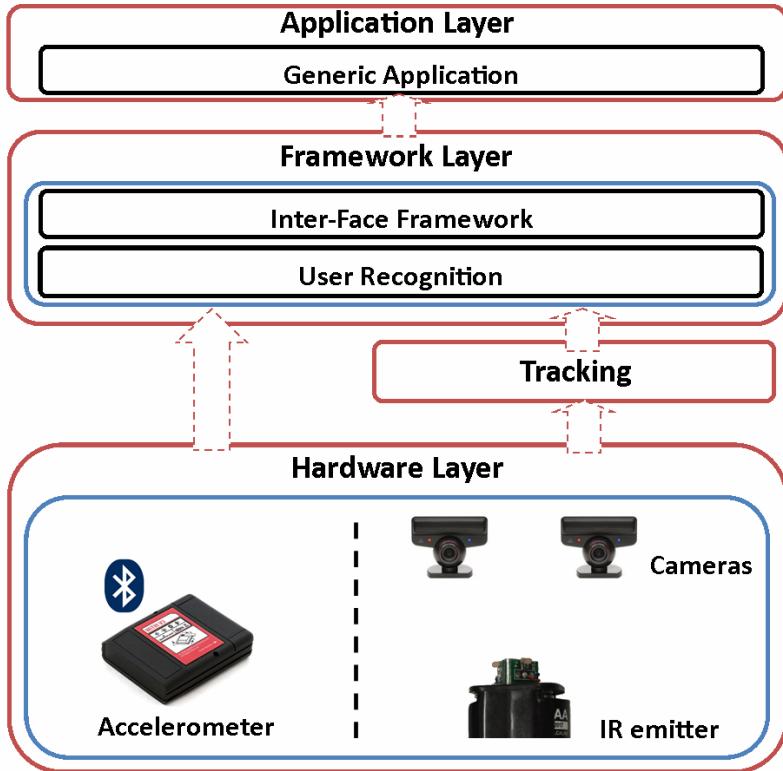


Fig. 1. PlusControl system architecture

4 Prototype

As shown in Fig. 1, the system has been integrated in an innovative framework, called Inter-Face that manages the creation of interactive surfaces supporting multi-user interaction [20]. Standard applications in a Windows environment, in fact, do not allow the use of multiple cursors. The Inter-Face Framework, instead, allows multiple inputs (i.e. several users interacting at the same time on the same surface) and combines multiple technologies (i.e. it integrates different sensing technologies such as acoustic, optic and RFID in order to support multimodal interaction). The PlusControl system has been integrated in the Inter-Face Framework as a new interaction modality.

The PlusControl system can recognize different gestures. Turning the left wrist counter-clockwise performs the press-click action. Turning the left wrist clockwise releases the click. While the press-click state is activated, turning the right wrist allows a rotation of 90° of the selected object in the application. This system allows configuring the devices symmetrically opposite to the previous explication; therefore it provides the possibility to perform gestures specularly. A quick horizontal movement in opposite directions of the hands allows zooming in or out the object.

The quick up-down gesture is used to recognize the user when the pointer is lost, exploiting the correlation of information from the tracked blob and the accelerometer of the arm in which the infrared emitter led is worn.

Two prototypes of the infrared emitter have been developed. The first one can be worn on a finger (Fig. 2 a)), while the other one can be held in the hand (Fig. 2 b)). Both prototypes have three IR LEDs placed on a triangular support; they allow tracking the IR source even with high inclination angles between the axis of the camera and the axis of the emitter. Both emitters are powered by two 1.5V AA batteries, which allow an operative time of about 48 hours. The emitters are tracked by two filtered Sony PS3 Eye cameras with a resolution of 640x480 pixels at 60 frames per second. Two Bluetooth SparkFun WiTilt V3.0 accelerometers are worn by the user on the wrists and they operate at a frequency of 50 Hz.



Fig. 2. Prototypes of the body worn infrared emitter (*left*) and of the handheld infrared emitter (*right*)

5 Testing Methodology

Some preliminary tests have been performed in order to assess the effectiveness and the usability of the system, also in a cooperative work scenario, by subjective and objective evaluations. We have invited 10 students (8 boys and 2 girls) to evaluate the system. We have created three applications to test this system and its many features.

The test was composed of 4 steps.

The step 1 consisted in solving a puzzle (a 1024x768 pixels image) utilizing in a first time the mouse and after our system. The puzzle was composed of 16 pieces (each one of 256x192 pixels) that were randomly distributed on a 4x4 table, as shown in Fig. 3 b). The users had to put the pieces in the right table cells to visualize the

original image. The screen was projected on a wall with a 1024x768 pixels resolution. We gave to the users the time to understand the image to compose and to try the control system. After this phase, we asked them to solve the puzzle and we recorded the time.

The step 2 was similar to the first one. The image to be recomposed had a dimension of 1280x1024 pixels and it was split in 16 pieces (each one of 320x256 pixels). The application was displayed with two monitors with resolution at 1680x1050 pixels. The interactive surface of the application was of 3280x1024 pixels. The 4x4 table was in the left side of the screen and the images randomly distributed in the right side, as shown in Fig. 3 a). The users had to put the pieces in the right table cells to visualize the original image. The users have utilized before the mouse and after the PlusControl system while we were recording the time.

The step 3 consists of solving the puzzle of the step 1 and of the step 2 in cooperation with another user. We recorded the time.

In the step 4 the users had to manipulate the images in an application using the gestures presented in the Architecture paragraph. The visual interface of this application is shown in Fig. 3 c).

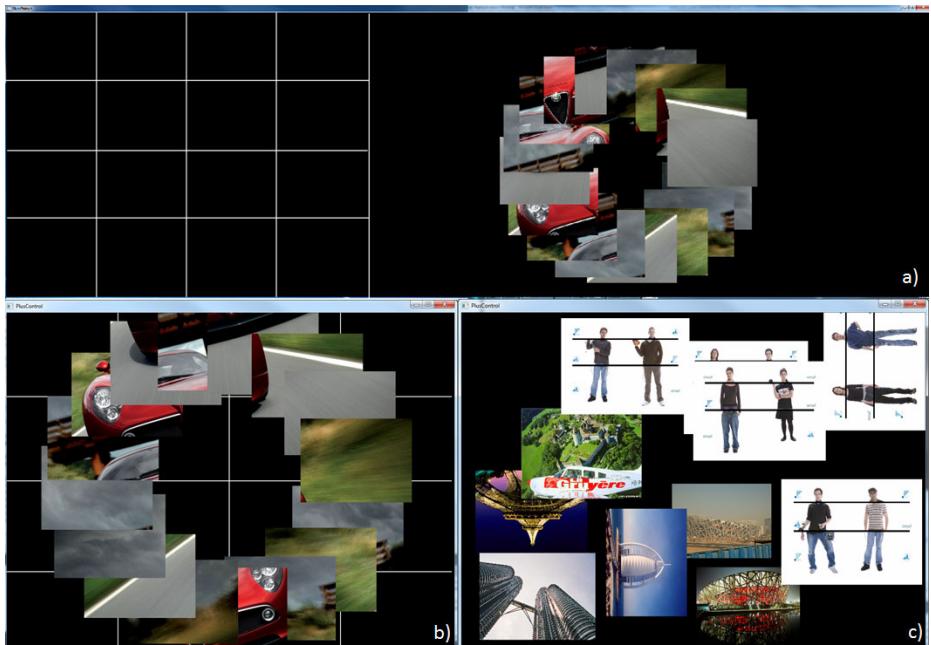


Fig. 3. a) The puzzle application for the test with the 3280x1024 px screen. b) The puzzle application for the test with the 1024x768 px screen. c) The application for the free manipulation of the images.

Finally, the participants had to fill in a questionnaire, rating the system features according to a 7-point Likert scale. The evaluation parameters were comfort, precision, learnability and cooperation. Comfort considers how tired the user was after the test

session and how easy was reaching all the parts of the screen. Precision is strictly related to the easiness and accuracy of the cursor and commands control. Learnability refers to how much the user, which utilized PlusControl for the first time, got acquainted in performing pointing action and gestures. Cooperation measures how much a participant has appreciated collaborate with another person and if he had difficulties to accomplish the given task. The questionnaire included a space for suggestions and comments.

6 Results

Tests have been executed following the methodology presented in Section 5. Results of these tests are discussed in this Section. Fig. 4 presents average solving time and relative standard deviation of steps 1, 2 and 3. In the graphic on the left are presented the results obtained with the 1024x768 pixels screen projected on the wall. The graphic on the right shows the results in the large displays scenario. Using the PlusControl system, the participants achieved a solving time that is 24% and 33% higher than utilizing the mouse, respectively to the step 1 and 2. Solving times have been drastically reduced in cooperation scenarios. Results of the subjective evaluation are presented in **Table 1**. Users expressed their appreciation about easiness of use and pointer reactivity. On the other hand they have suggested improving the reactivity of the gesture recognition subsystem.

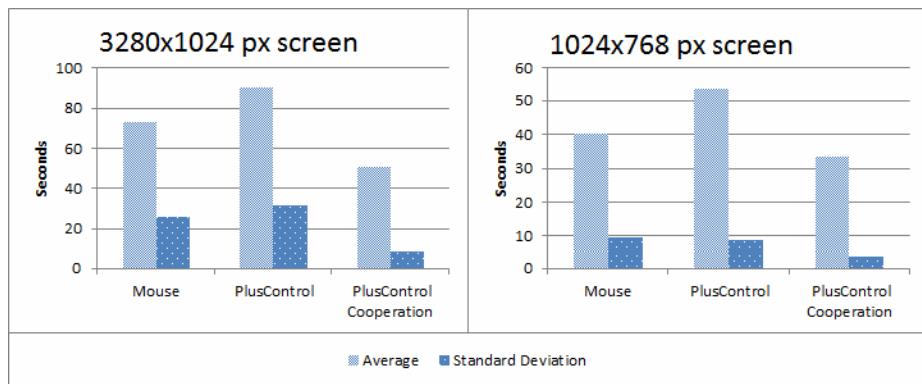


Fig. 4. Average time (and standard deviation) for puzzle completion on 3280x1024 px screen (*left*) and 1024x768 px screen (*right*)

Table 1. Results of the subjective evaluation of PlusControl main features according to a 7-point Likert scale

Features	Average	Standard Deviation
Comfort	5.7	0.7
Precision	5.5	0.8
Learnability	5.3	0.7
Cooperation	6.2	0.6

7 Conclusions and Future Works

In this article a novel interaction system that allows to several users to cooperate on a shared large screen has been presented. PlusControl combines an infrared pointing system with the recognition of several natural gestures using accelerometers. The test showed that users appreciated the system and that it can be very useful in cooperative scenarios. In the future works the system will be tested with more than two users at the same time, in order to infer the maximum number of users working contemporaneously according to technology constraints. Moreover we will improve the performances of the system by developing a specific device which will integrate the triaxial accelerometer and the infrared emitter used so far, plus a gyroscope.

References

1. Bolt, R.A.: “Put-that-there”: Voice and Gesture at the Graphics Interface. ACM SIGGRAPH Computer Graphics 14, 262–270 (1980)
2. Krahnstoever, N., Kettebekov, S., Yeasin, M., Sharma, R.: A real-time framework for natural multimodal interaction with large screen displays. IEEE Comput. Soc., Los Alamitos (2002)
3. Kela, J., Korppiä, P., Mäntylä, J., Kallio, S., Savino, G., Jozzo, L., Marca, S.D.: Accelerometer-based gesture control for a design environment. Personal and Ubiquitous Computing 10, 285–299 (2005)
4. Amershi, S., Morris, M.R., Moraveji, N., Balakrishnan, R., Toyama, K.: Multiple mouse text entry for single-display groupware. ACM Press, New York (2010)
5. Yin, Y., Davis, R.: Toward Natural Interaction in the Real World: Real-time Gesture Recognition. ACM Press, New York (2010)
6. Shintani, K., Mashita, T., Kiyokawa, K., Takemura, H.: Evaluation of a Pointing Interface for a Large Screen Based on Regression Model with Image Features. IEEE, Los Alamitos (2010)
7. Baudisch, P.: Interacting with Large Displays. Computer 39, 96–97 (2006)
8. Morrison, G.D.: A camera-based input device for large interactive displays. IEEE Computer Graphics and Applications 25, 52–57 (2005)
9. Fukuchi, K.: A laser pointer/Laser trails tracking system for visual performance. In: Costabile, M.F., Paternó, F. (eds.) INTERACT 2005. LNCS, vol. 3585, pp. 1050–1053. Springer, Heidelberg (2005)
10. Zhang, L., Shi, Y., Chen, B.: NALP: Navigating Assistant for Large Display Presentation Using Laser Pointer. In: First International Conference on Advances in Computer-Human Interaction, pp. 39–44 (2008)
11. Davis, J., Chen, X.: Lumipoint: multi-user laser-based interaction on large tiled displays. Displays 23, 205–211 (2002)
12. Nakamura, T., Takahashi, S., Tanaka, J.: Double-crossing: A new interaction technique for hand gesture interfaces. In: Lee, S., Choo, H., Ha, S., Shin, I.C. (eds.) APCHI 2008. LNCS, vol. 5068, pp. 292–300. Springer, Heidelberg (2008)
13. Oh, J.Y., Stuerzlinger, W.: Laser pointers as collaborative pointing devices. Graphics Interface, pp. 141–150, Citeseer (2002)
14. Jiang, H., Ofek, E., Moraveji, N., Shi, Y.: Direct pointer. ACM Press, New York (2006)

15. Barbieri, R., Farella, E., Benini, L., Ricco, B., Acquaviva, A.: A low-power motion capture system with integrated accelerometers gesture recognition applications. In: 2003 International Symposium on VLSI Technology, Systems and Applications. Proceedings of Technical Papers (IEEE Cat. No.03TH8672), pp. 418–423 (2004)
16. Liu, J., Wang, Z., Zhong, L., Wickramasuriya, J., Vasudevan, V.: uWave: Accelerometer-based personalized gesture recognition and its applications. In: IEEE International Conference on Pervasive Computing and Communications, pp. 1–9 (2009)
17. Mäntylä, J., Kela, J., Korpipää, P., Kallio, S.: Enabling fast and effortless customisation in accelerometer based gesture interaction. In: Proceedings of the 3rd International Conference on Mobile and Ubiquitous Multimedia - MUM 2004, pp. 25–31 (2004)
18. Fu, L.C., Fellow, I.: Gesture stroke recognition using computer vision and linear accelerometer. In: 8th IEEE International Conference on Automatic Face & Gesture Recognition, pp. 1–6 (2008)
19. Schlömer, T., Poppinga, B., Henze, N., Boll, S.: Gesture recognition with a Wii controller. In: Proceedings of the 2nd International Conference on Tangible and Embedded Interaction - TEI 2008 (November 2008)
20. Mugellini, E., Abou Khaled, O., Pierroz, S., Carrino, S., Chabbi Drissi, H.: Generic framework for transforming everyday objects into interactive surfaces. In: Jacko, J.A. (ed.) HCI International 2009. LNCS, vol. 5612, pp. 473–482. Springer, Heidelberg (2009)

Gesture Identification Based on Zone Entry and Axis Crossing

Ryosuke Aoki¹, Yutaka Karatsu², Masayuki Ihara¹, Atsuhiko Maeda¹,
Minoru Kobayashi¹, and Shingo Kagami³

¹ NTT Cyber Solutions Laboratories, NTT Corporation,

1-1 Hikari-no-oka, Yokosuka-Shi, Kanagawa, 239-0847 Japan

² Graduate School of Media and Governance, Keio University,

5322 Endou, Fujisawa, Kanagawa, 252-0882 Japan

³ Graduate School of Information Sciences, Tohoku University,

6-6-01 Aramaki Aza Aoba, Aoba-ku, Sendai, 980-8579 Japan

{aoki.ryosuke,ihara.masayuki,maeda.atsuhiko,

kobayashi.minoru}@lab.ntt.co.jp,

karasu@ht.sfc.keio.ac.jp,

swk@ic.is.tohoku.ac.jp

Abstract. Hand gesture interfaces have been proposed as an alternative to the remote controller, and products with such interfaces have appeared in the market. We propose the vision-based unicursal gesture interface (VUGI) as an extension of our unicursal gesture interface (UGI) for TV remotes with touchpads. Since UGI allows users to select an item on a hierarchical menu comfortably, it is expected that VUGI will yield easy-to-use hierarchical menu selection. Moreover, gestures in the air such as VUGI offer an interface area that is larger than that provided by touchpads. Unfortunately, since the user loses track of his/her finger position, it is not easy to input commands continuously using VUGI. To solve this problem, we propose the dynamic detection zone and the detection axes. An experiment confirms that subjects can input VUGI commands continuously.

1 Introduction

Large TVs have been widely adopted and users are demanding a comfortable input method to utilize services such as on demand TV and EPG through these large displays. These existing services employ hierarchical menu selection for content handling. In addition, hierarchical menu selection may be used to select content for interactive digital signage. Toward easy-to-use hierarchical menu selection, it is better that users are able to select the target content by inputting small commands with the least cognitive and physical burdens. Hand gesture interfaces have been proposed as an alternative to the remote controller, and products with such interfaces have appeared in the market [1]. Our recent proposal is the unicursal gesture interface (UGI), which allows easy-to-use hierarchical menu selection, see Fig. 1 [2].

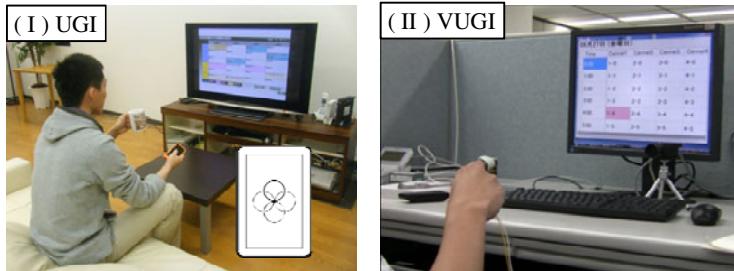


Fig. 1. (I) Unicursal gesture interface (II) Vision-based unicursal gesture interface

In this paper, we extend UGI, which is based on 2D gesture, to Vision-based UGI (VUGI), which employs 3D gestures. VUGI uses the forefinger instead of the thumb and a camera instead of a touchpad to detect finger position, see Fig. 1. VUGI pairs a unicursal figure with orientation to create a command in the same way as UGI. However, the cameras permit the area in which gestures can be captured to be basically unlimited, unlike UGI whose operation area is limited to the touchpad provided by the handheld device. The freedom provided by VUGI allows the user to lose track of his/her finger position, so it is not easy to input commands continuously via VUGI.

To solve this problem, we propose a gesture recognition method for VUGI that uses our dynamic detection zone and detection axes. An experiment confirms that users can input commands continuously by the proposed method.

In the following sections, we will review related works to explain the value of UGI. We evaluate the influence of size of detection zone to elucidate the problems posed by VUGI. We discuss the design of the dynamic detection zone and the dynamic axes to solve the problems. We evaluate the performance of the proposed method. Finally we conclude by discussing VUGI performance and future work.

2 Related Work

We explain the value of UGI by examining the suitability of existing hand gestures for interface control. Existing gesture interfaces fall into three categories: A) recognition of hand form, B) recognition of absolute hand position and/or absolute hand orientation, and C) recognition of the shape or path traced by the hand.

Gesture recognition systems accept a command upon recognizing a hand form such as OK sign or V sign, common in nonverbal communication [3-5]. These gestures are used to execute functions such as keyboard shortcut. However, these gestures are not suitable for smoothly inputting many commands needed to select target content such as hierarchical menu selection because of wasted motion from a hand form to next hand form.

Gesture recognition systems have been described that associate an item on a menu according to absolute hand position or absolute hand orientation as measured by sensors [6-7]. To select items precisely, the areas for each item must be large. Since

the overall operation area is limited by sensor performance or human anatomy, the number of items on the menu is limited. When users make use of services such as EPG and on demand TV, both of which contain a lot of content, users must switch between menus so frequently that their physical burden is large.

Interface commands can be based on the shape or path drawn by the hand or finger [8-13]. One critical problem with gesture recognition is to detect the start and end points of gestures made with continuous motion. For example, gesture starts/(ends) when user's arm is/(isn't) stretched to a screen [1] or when user's finger push/(lift from) the physical key [14]. However, the gestures include wasted motion from end of gesture to start of next gesture. Since it is user's physical burden, it is not easy to input commands continuously such as hierarchical menu selection by these gestures.

UGI is a thumb-gesture interface for handheld devices with touchpads [2]. UGI pairs a shape with orientation to create a command. The shapes are simple unicursal figures whose paths start and end at the same position and that are easy for the user's thumb to remember. The orientation is grounded on the start position of the gesture and is simple. The features of the method are as follows. (1) Users can use simple finger actions within small spatial extent to create many TV commands. (2) Users can directly access different layers of the hierarchical menu via simple to remember commands as well as moving rapidly across each layer for item selection by changing the unicursal shape. (3) Time is not wasted in moving the finger to the start position of the next gesture. (4) The simple figures are used so often in daily life that users find it easy to remember the UGI gestures. Therefore, UGI provides users with a command-rich interface that imposes minimal physical and cognitive loads and allows users to select target content by inputting small commands. In the next section, we introduce the vision-based unicursal gesture interface as an extension of UGI.

3 Vision-Based Unicursal Gesture Interface (VUGI)

VUGI is a forefinger gesture interface and the finger position is detected by vision-based image processing. VUGI pairs a shape with orientation to create commands. We use simple unicursal figures whose paths start and end at the same position and are easy for the user's forefinger to remember and recreate, see Fig. 2(I). As one example, the unicursal line shown in Fig. 2(I)(c) is drawn by the forefinger moving from point "A" to point "B" and returning to point "A". The orientation is grounded on the start position of the gesture. For example, see Fig. 2(II). This interface can be activated in several ways such as pushing a physical key attached to a remote controller, hand-waving gesture, or motion of extending one's arm to TV screen [1]. In this paper, the gesture recognition system is activated when the user's forefinger is positioned at the center of the camera image.

The usability of this method depends on reliably detecting gesture start and end points. Since pixel-wise detection is impractical, this method introduces a detection zone, the circular zone shown in Fig. 2(III). A gesture starts when the forefinger leaves the detection zone and ends when the forefinger returns to the detection zone. A key parameter is the size of the detection zone.

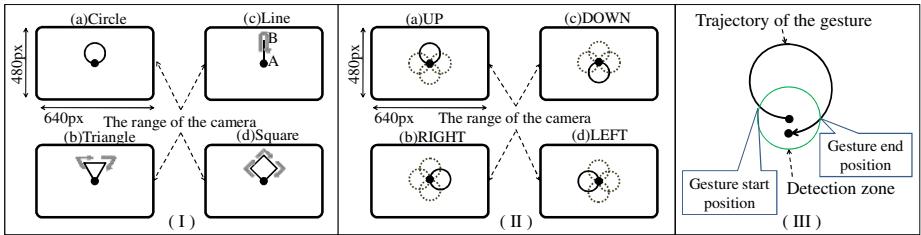


Fig. 2. (I) Simple unicursal figures (II) Gesture orientation of the unicursal figures (III) Gesture detection by Detection zone

4 Influence of Size of Detection Zone

We examined the degree to which start and end points coincided forefinger and impact of detection zone size on VUGI usability.

4.1 Design of Experiment

A within-subject experiment was conducted. The independent variable was radius of the detection zone (15px, 30px). The dependent variables were the number of gestures recognized and distribution of the positions at which the subject's forefinger returned to the detection zone and the positions at which it stopped after the gesture was recognized. The subject was a 22-year old man.

4.2 Apparatus

The experiment used a 3.16 GHz Intel Xeon CPU running Windows Vista. The display, connected to the PC via VGA cable, was a 20.1-inch LCD (EIZO FlexScan S2000) with 1280 * 1024 pixel resolution. The gesture capture camera, connected to the PC via IEEE-1394BUS, was a Grasshopper GRAS-03K2M supplied by Point Grey Research. Sampling frequency of the camera was 30 fps where each gray scale image contained 640 * 480 pixels. Subject's forefinger was in front of the camera. A white light-emitting diode was attached to the subject's forefinger to detect position on the camera image. The distance between the subject's forefinger and the camera was 60 cm. Fig. 3(I) shows the experimental setup.

4.3 Procedure

The subject sat on a chair, his elbow rested on an arm of the chair and his forefinger was pointed at the camera, see Fig. 3(I). He moved his forefinger while looking at the display. The experiment flow is shown below. A window with 640 * 480 pixel resolution was displayed on the screen when starting the application. The subject watched the images taken by the camera in real time through the window. The subject initialized the detection zone by pointing to the center of the camera image. The

subject drew an “up” circle and a message was displayed on the screen when his forefinger returned to the detection zone, which remained fixed at the center of the screen. The subject stopped forefinger movement after watching the message. The subject repeated the input operation 150 times.

4.4 Results and Discussion

Fig. 3(II) shows the positions at which the subject’s forefinger returned to the detection zone and the positions at which it stopped after the gesture was recognized. Fig. 3(III) shows the trajectory of one “up” circle drawn by the subject. Fig. 4(I) shows the number of gestures recognized for the 150 inputs.

Fig. 3(II) indicates that at end of one gesture the finger tends to overshoot the center of the zone and stop around the far-side boundary. Since the finger oscillates when the finger stops, this oscillation yields erroneous recognition results (Fig. 3(III)). To solve this problem, we change the detection zone dynamically (section V).

Using a detection zone radius of 15px yielded a lower gesture recognition rate than a radius of 30px. This is because it was not easy for the subject’s forefinger to return to the 15px detection zone shown in Fig. 4(II). However, the finger is not far from the initial position during inputting commands continuously since the smaller radius yielded a tighter grouping of end positions. In Subsequent experiment we assumed a default detection zone radius of 15px. We propose a way to solve the problem of the forefinger failing to return to the detection zone.

5 Dynamic Detection Zone and Detection Axis Technique

To overcome the erroneous recognition caused by finger oscillation, we propose the dynamic detection zone; its size changes according to forefinger position, see Fig. 5(I). The detection zone is initially large. When the forefinger leaves the initial detection zone, it is replaced by a smaller detection zone. Finally, after the forefinger enters the small detection zone and the gesture is recognized, the large detection zone is reset.

It is impossible to continue gesture input until the forefinger clearly re-enters the detection zone. To solve this problem, we propose the detection axis technique. Vertical and horizontal detection axes are centered on the initial position of the gesture. Each axis is triggered when the forefinger enters one of the gray zones shown in Fig. 5(II). When the forefinger crosses the horizontal (vertical) axis for the second time when making a vertical (horizontal) unicursal figure (Fig. 5(III) top and bottom, respectively), the method detects that the gesture has been completed. The detection zone is then re-centered on the point at which the finger crossed the detection axis, see Fig. 6(I). This implies that the detection zone can wander slightly. This is not a problem since the detection zone will come close to the initial position which the detection zone is made after activating the gesture recognition system when the forefinger enters the detection zone, see Fig. 6(II).

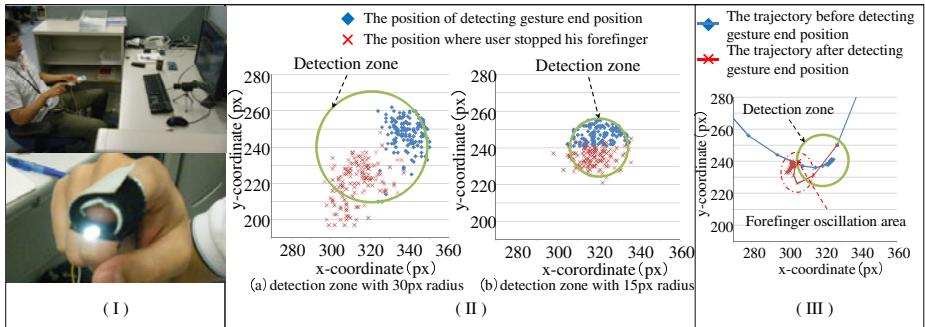


Fig. 3. (I) Experiment environment (II) The positions at which the subject's forefinger returned to the detection zone and the positions at which it stopped after the gesture was recognized. (III) The trajectory of one "up" circle drawn by the subject.

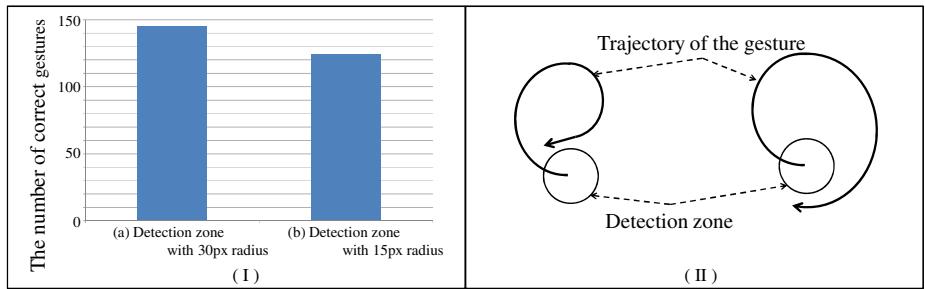


Fig. 4. (I) the number of gestures recognized for the 150 inputs (II) The problem of the forefinger failing to return to the detection zone

6 Evaluation of Dynamic Detection Zone

We examined the impact of the dynamic detection zone technique on the number and kinds of error.

6.1 Design of Experiment

A within-subject experiment was conducted. The independent variable was the radius of the large detection zone (15px, 30px, 45px, 60px, 120px). Radius of the small detection zone was fixed at 15px. The dependent variables were the number of gesture recognition errors and the kinds of error. Each trial finished when 50 gestures had been recognized.

The subject, experimental setup, and apparatus were the same as the experiment in Section 4.

6.2 Results and Discussion

Fig. 7(I) shows the two main kinds of errors. One is that the subject's forefinger oscillated around the boundary of the large detection zone. The other is that the

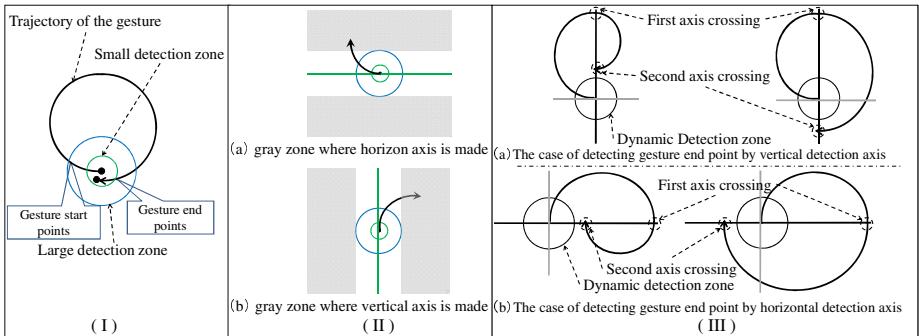


Fig. 5. (I) Gesture detection by the dynamic detection zone (II) Gray zone where detection axis made (III) Gesture detection by the detection axis

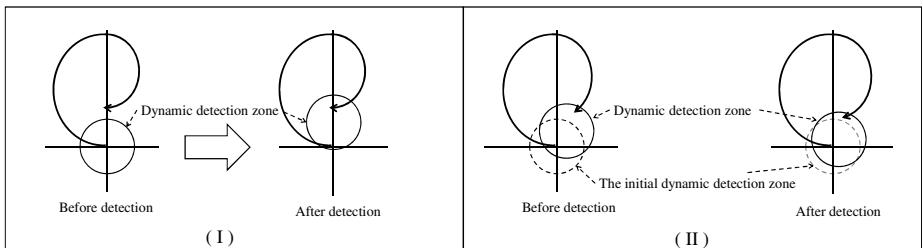


Fig. 6. (I) Moving the detection zone after detection axis detects (II) Moving the detection zone after detection zone detects

gesture was not recognized since the subject's forefinger did not return to the small detection zone. Fig. 7(II) shows the number of errors recorded in achieving the successful recognition of 50 gestures. Setting large detection zone radius at 45px yielded the fewest errors.

Any large detection zone radii greater than 15px eliminated the occurrence of oscillation error. However, radii greater than 45px increased the number of errors since the subject's forefinger did not leave the large detection zone, see Fig. 7(III).

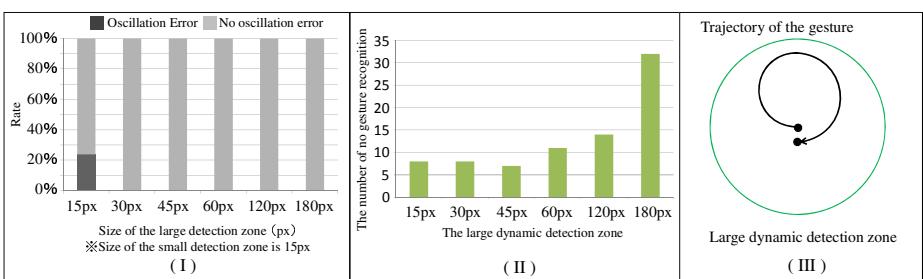


Fig. 7. (I) The kinds of error (II) The number of errors recorded in achieving the successful recognition of 50 gestures (III) The error when the large detection zone is large

7 Evaluation of Detection Axes

We examined the continuous input of commands when both proposed techniques were used: the dynamic detection zone and detection axes.

7.1 Design of Experiment

A within-subject experiment was conducted. The independent variables were Method (Only dynamic detection zone and the combination of dynamic detection zone and detection axes). No subject was permitted to look at forefinger position after setting the forefinger initial position. The dependent variable was the number of gestures recognized continuously.

Six unicursal gestures were used to operate an EPG on a screen, see Fig. 8(I) and Fig. 8(II). Up/Down circle moves the cursor on the menu above/below. As well, Left/Right circle moves the cursor on the menu to the left/right. Up/Down unicursal line accesses the EPG page of yesterday/tomorrow. Since we assume that users operate the EPG searching content, users stop gesture per one command and input commands continuously.

The experimental setup is the same as in Sections 3 and 5.

7.2 Subject

Four subjects (one woman) were conducted. All were right handed. Ages ranged from 22 to 31 years.

7.3 Procedure

First, all the subjects practiced the gestures used in this experiment. They were informed when their forefinger reached the initial position which was displayed on the center of the camera image on the screen. After their forefinger reached the initial position, the 6 gestures were displayed individually on the screen, and each subject traced the shape by their forefinger.

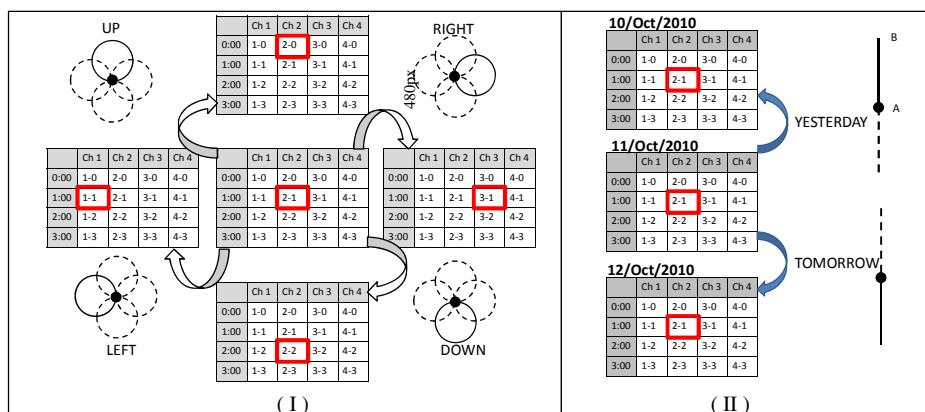


Fig. 8. An example of EPG operation

Next, all subjects practiced the task demanded by this experiment. Each subject input the figure indicated by text on the screen of the PC. The command “yesterday” corresponded to the up unicursal line, “tomorrow” to down unicursal line, “up” to up circle, “down” to down circle, “left” to left circle, and “right” to right circle. The subjects practiced these VUGI commands for ten minutes.

Finally, the subjects performed 3 trials. Each trial finished when 50 commands had been input continuously or was interrupted when each subject lost the position of the detection zone or the detection axes and couldn't input a command.

7.4 Result and Discussion

Fig. 9(I) shows the mean number of commands input continuously for each subject and all subjects. A t-test showed that the number of commands input continuously using the dynamic detection zone and detection axes was significantly larger than that achieved by using only the dynamic detection zone ($F(1,7)=10.2443$, $p<0.05$). We focus on two subjects who failed to achieve the input of 50 commands continuously. Fig. 9(II) shows the number of commands input continuously in each trial. The input numbers reached 50 in the second trial and third trial. That is, all subjects could input 50 commands continuously except in the first trial.

Fig. 9(III) shows that the mean recognition rate was 72% for the 6 gestures. This result suggests the need for further improvement in the recognition process.

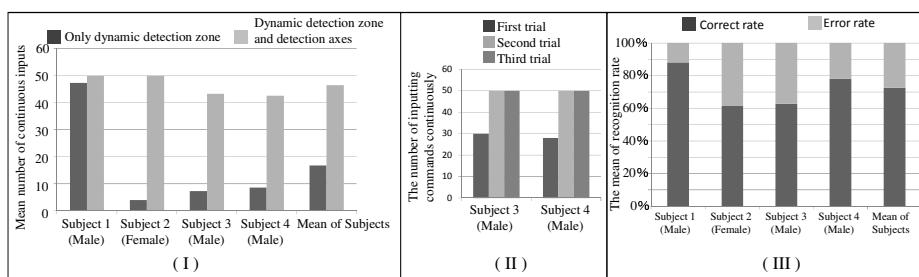


Fig. 9. (I) (II) the mean number of commands input continuously (III) The mean recognition rate

8 Conclusion

We proposed the dynamic detection zone and the detection axis technique to input commands continuously via vision-based unicursal gesture interface. Experiments confirmed that users could input commands continuously. We intend to improve the recognition rate of gestures.

References

- Shao, L., Shan, C., Luo, J., Etoh, M.: *Multimedia Interaction and Intelligent User Interfaces: Principles, Methods and Applications (Advances in Pattern Recognition)*, pp. 107–128. Springer, Heidelberg (2001)
- Aoki, R., Ihara, M., Maeda, A., Watanabe, T., Kobayashi, M.: Unicursal Gesture Interface for TV Remote with touch screen. In: ICCE 2011, pp. 101–102 (2011)

3. Gupta, L., Ma, S.: Gesture-Based Interaction and Communication: Automated Classification of Hand Gesture Contours. *IEEE Transactions on Applications and Reviews* 31(1), 114–120 (2001)
4. Stern, H.I., Wachs, J.P., Edan, Y.: Hand gesture vocabulary design: a multicriteria optimization. In: *IEEE International Conference on Systems, Man and Cybernetics*, vol. 1, pp. 19–23 (2001)
5. Althoff, F., Lindl, R., Walchshausl, L.: Robust multimodal hand- and head gesture recognition for controlling automotive infotainment systems. *VDI-Tagung – Der Fahrer im 21. Jahrhundert*, Braunschweig, Germany (2005)
6. Ni, T., McMahan, R.P., Bowman, D.A.: Tech-note: rapMenu: Remote Menu Selection Using Freehand Gestural Input. In: *IEEE Symposium on 3DUI 2008*, pp. 55–58 (2008)
7. Aoki, R., Maeda, A., Watanabe, T., Kobayashi, M., Abe, M.: Twist&Tap: Text Entry for TV Remotes Using Easy-to-Learn Wrist Motion and Key Operation. *IEEE Transaction on Consumer Electronics*, 161–168 (2010)
8. Alon, J., Athitsos, V., Sclaroff, S.: Accurate and Efficient Gesture Spotting via Pruning and Subgesture Reasoning. In: *Proc. IEEE ICCV Workshop Human Computer Interaction*, pp. 189–198 (2005)
9. Lee, H., Kim, J.: An HMM-Based Threshold Model Approach for Gesture Recognition. *IEEE Transaction Pattern Analysis and Machine Intelligence* 21(10), 961–973 (1999)
10. Morguet, P., Lang, M.: Spotting Dynamic Hand Gestures in Video Image Sequences Using Hidden Markov Models. In: *Proc. IEEE Int'l Conf. Image Processing*, pp. 193–197 (1998)
11. Oka, R.: Spotting Method for Classification of Real World Data. *The Computer Journal* 41(8), 559–565 (1998)
12. Zhu, Y., Xu, G., Kriegman, D.: A Real-Time Approach to the Spotting, Representation, and Recognition of Hand Gestures for Human-Computer Interaction. *Computer Vision and Image Understanding* 85(3), 189–208 (2002)
13. Perlin, K.: Quikwriting: continuous stylus-based text entry. In: *UIST 1998* (1998)
14. Steven, J., Castellucci, Mackenzie, I.S.: UniGest: Text Entry Using Three Degrees of Motion. In: *CHI 2008*, pp. 3549–3554 (2008)

Attentive User Interface for Interaction within Virtual Reality Environments Based on Gaze Analysis

Florin Barbuceanu¹, Csaba Antonya¹, Mihai Duguleana¹, and Zoltan Rusak²

¹ Transilvania University of Brasov, Department of Product Design and Robotics,
29 Eroilor Blvd., Brasov, Romania

² Technical University of Delft, Department of Design Engineering
Landbergstraat 15, 2628CE, Delft, The Netherlands
{Florin.Barbuceanu, Antonya, Mihai.Duguleana}@unitbv.ro,
z.rusak@tudelft.nl

Abstract. Eye movements can carry a rich set of information about someone's intentions. In the case of physically impaired people gaze can be the only communication channel they can use. People with severe disabilities are usually assisted by helpers during everyday life activity, which in time can lead to a development of an effective visual communication protocol between helper and disabled. This protocol allows them to communicate at some extent only by glancing one towards the other. Starting from this premise, we propose a new model of attentive user interface featured with some of the visual comprehension abilities of a human helper. The purpose of this user interface is to be able to identify user's intentions, and so to assist him/her in the process of achieving simple interaction goals (i.e. object selection, task selection). Implementation of this attentive interface is accomplished by way of statistical analysis of user's gaze data, based on a hidden Markov model.

Keywords: gaze tracking, eye tracking, attentive user interface, hidden Markov model, disabled people.

1 Introduction

Special eye movements (i.e. blinking, rough saccadic shifts or fixations) can be easily recognized and associated with particular intentions or emotional states [1]. Fast blinking could signal the presence of stress, while looking around randomly might indicate a lack of interest. Meaningful information can be extracted from eye movements and interpreted through special attentive user interfaces that are designed to detect the relation between eye movements and user's intentions. When, for example, an interlocutor gazes longer towards another, it is supposed he/she is expecting an answer [2]. Also, previous work shows that by using a virtual assistant a social natural communication protocol based on eye movements can be induced to the user [3]. In this case users can gaze directly towards assistant's face when they want to capture its attention, then towards a certain object to indicate the interest on it, and eventually back to the agent, waiting for a response. This small set of procedures is similar to a simple human to human visual communication protocol. Based on these

results, a new generation of virtual assistants has been implemented [4], which follows the gaze of the user and adapts the content displayed on a screen depending on subject's interests. This type of user interfaces constitutes a powerful tool for the disabled, reducing the workload of the interaction procedures.

1.1 Previous Work on Attentive User Interfaces

In a general context, the state of uncertainty is characterized by random or irregular eye movements. When all the other communication means are altered, eye movements and gaze gestures can successfully be used to exchange valuable data with the disabled. An attentive interaction interface for disabled should be able to copy some of the comprehension abilities of a human helper and assist him/her when the helper is not around. In order to achieve this, the interface must track the gaze of the users and infer their intentions based on the contextual information of the environment and the relation between gazed objects [5]. A comparative study with the aim of evaluating user's contextual preferences and states of uncertainty based on the analysis of eye movements is presented in [6]. iDict is an application designed to ease the translation of unknown words from a text, through an attentive user interface able to determine when the users are in a state of uncertainty. It compares the measured data about eye movements to existing data on normal dynamics of eye movements measured during text reading. This approach has been extensively studied in the literature. In iDict the process of eye movement analysis is triggered, if at any given time eye movements are abnormal from the way they are expected to be, a process of eye movement analysis is triggered. The results of the eye movement's analysis could for instance state that the process of text comprehension is discontinued and the attentive interface can or cannot take several alternative actions to help the reader getting back on track. One possible action in this context is to translate the words where the discontinuity occurred, and check whether this helps the users going further with their reading. This feature enables users to speed up the process of text comprehension, since normally they would have been constrained to open a dictionary to search for the unknown word. Using an algorithm based on a comprehension difficulty factor [7], an amazing 91% success rate in detecting unknown words, and a very small 2.4% false alarm rate were achieved.

A second study evaluates the uncertainty during the process of reviewing answers given previously to a set of questions. Based on eyes movements' analysis, a "strength-of-belief" (SOB) factor has been defined to characterize the level of uncertainty. In this study scanning eye movements are considered correspondent to high rates of SOB factor, while transition movements indicate low rates. The users have also given a subjective estimation of SOB for each answered reviewed, which has been compared with the real value of SOB. The results indicate that user's estimations of answer correctness are closely related to the values of the SOB factor determined from eye movements' analysis. Their technique can thus successfully be applied for detection of user's uncertainty.

AutoSelect is a user preference detection system based on eye movements' analysis, exploiting the so called gaze "cascade effect" discovered in [8]. This effect consists in the gradually gaze shifting between two similar objects, with the tendency to lean more towards the preferred objects, by the time the preference clears up in

user's mind. The success rate achieved by the AutoSelect system in an experiment based on 8 subjects, 4 males and 4 females is 81%,

The iTourist system provides interactive information about the location of interest on a city map, based on eye movements' analysis [9]. It was designed with the purpose of testing how much extra information can be provided to the user based only on gaze tracking. Every time the users gaze longer over a certain location on the map the system displays guiding information about it. The results indicate a clear advantage over a human guide, because it is hard for humans to constantly follow users' gaze and figure out what they are interested in, while iTourist can easily do this, constantly providing guiding information about the locations of interest.

2 Conscious vs. Unconscious Action Triggering

Face-to-face communication between humans is a complex process of information transfer both verbally and visually. Not only the words count for the data exchanges, but also the tone of the voice, speech rhythm, face expression, all these important clues about the intrinsic emotional and cognitive states. The results in [10] shows that eyes and mouth are the dominant areas gazed during a dialog.

Simple eye gestures drive the communication within the Eye Bed interface research project [11]. A selection is made through a prolonged blink. Gazing around randomly signifies lack of interest, while staring indicates attention. If eyes get closed after lying in a bed, it means the user wants to sleep. Eyes opened after a long time sleep means the user will wake up, and so the light in the room can be opened, the radio may be started up etc. This kind of simple eye movements can be easily interpreted and implemented in effective attentive user interfaces.

Predefined conscious eye's gestures can carry cognitive data about user's intentions, significantly improving the communication between humans. They simplify the process of message extraction from eye movements data, but they also constrain the user to perform unnatural eye movements, overloading thus with motor control tasks the perceptual function of the visual channel [12]. In this case eyes play a role of an output communication channel, used to express user's intentions through consciously controlled eye movements. But because the eye is a perceptual organ not meant for motor controlling tasks [7], interaction actions based on gaze should be triggered without any tedious eye gestures, but rather they should be detected from the context. Specific sequences of eye movements can be sufficient to decide when an action needs to be triggered. An assistive interface able to detect the correct significance of a specific sequence of eye movements and take the corresponding action at the right time can free the user from performing unnecessary burden eye gestures. This means that the cognitive load at the user is transferred to the attentive interface, which needs to come with more sophisticated and more precise intention detection algorithms. When designing an attentive interface based on unconscious action triggering, there are several issues to confront with. For instance, the "Midas touch" problem appears when a false alarm dwell time selection is triggered [13]. The consequence of improper intention detection could reside in ineffective communication leading to frustration and eventually rejection. To compare, conscious

action triggering can improve the accuracy of the intention recognition through predefined eye gestures, at the cost of overloading the natural functions of the eyes, while unconscious action triggering is more natural, does not disturb the normal functions of eyes, with a possible cost of lower accuracy and higher complexity attentive systems implementations. If the accuracy problems can be minimized through highly robust attentive systems, then unconscious action triggering interaction would ultimately be the right interaction interface to be implemented. Just imagine how difficult the usability of the iDict application [6] would be if instead of automatic detection of uncertainty, the users should blink their eyes every time a translation of a word is needed.

3 The Experimental Setup

Our experiments were conducted in a virtual reality environment to test the usability of the human-computer interaction interfaces developed for people with severe locomotion disabilities. The virtual environment is a kitchen with several objects placed on the tabletop. The user can interact with these objects by looking towards them using a simple command based gaze interaction protocol, or a more complex attentive user interface based on a hidden Markov model implementation. Images are projected on a 3D stereoscopic visualization screen and the subject is immersed in the virtual environment through a pair of polarized 3D glasses. The eye tracking device used was the head-mounted model ASL H6-HS-BN [14].

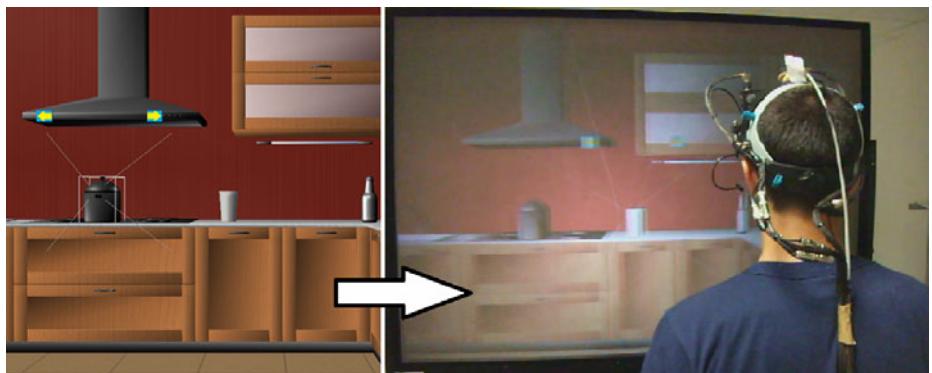


Fig. 1. A virtual kitchen containing objects the user can interact with: a bottle, a glass, and a pot. The virtual scene is projected on a 3D visualization screen in front of the user.

The graphics of the virtual environment is rendered by the XVR (Extreme Virtual Reality) software platform. The 3d model of the kitchen (Fig. 1) is composed by 3dsmax objects available at [15]. Selection of virtual objects available in the virtual kitchen is made by intersecting the line of sight, detected by the eye tracking device, and the objects, through the `IsColliding()` function, available whithin XVR SDK. Provided that the visual angle accuracy of the eye tracking device is 0.5° [14], using a

bar object with a given thickness, instead of a thin line, the chances of successful selection of objects through gaze is increased [16]. The distance between neighboring virtual objects is about 50 cm, and the distance to the user is 3 m. This corresponds to a spatial separation between two virtual objects of approximately 9^0 from the user's point of view (Fig.2).

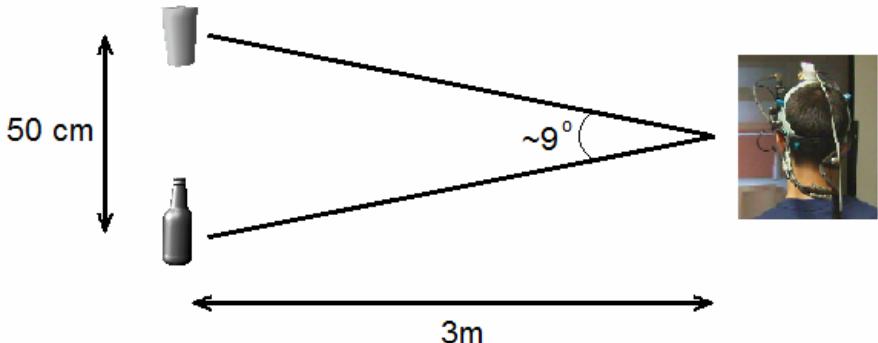


Fig. 2. Spatial separation of neighboring objects, from user's point of view

With the given eye tracking measurement accuracy of 0.5^0 , at 3m distance from the user, the accuracy of selection is about ± 2.6 cm around the real gaze point. The glass and the bottle at this distance have dimensions that can negatively influence the precision of eye tracking, so the selection can easily fail. In order to increase the success rate of selections, the thickness of the bar was chosen to be 10 cm. After some tests this thickness proved to generate most successful selections and it was adopted for further experiments.

4 Attentive User Interface Based on Hidden Markov Chains

Thoughts, intentions or preferences of a person are hidden deep inside the mind. Physically impaired persons, who cannot use conventional communication channels, face a major drama of not being able to express their thoughts or needs when necessary. Because of the hidden nature of these cognitive states, one solution for guessing them is to use a hidden Markov chains approach in the implementation of an attentive user interface based on gaze tracking.

A hidden Markov model is a hidden stochastic process which cannot be directly observable, but it generates dependent non-hidden stochastic emissions [17]. By analyzing a sequence of input data and knowing the transition and emission matrixes, it is possible to estimate the state they were generated from. In this implementation, the sequence of emissions consists in the recorded identification (ID) number of the objects gazed by the user over time (Tab. 1). Several objects have been displayed in a virtual kitchen, a bottle, a glass, a pot and a tap. Each object has an ID of its own, as follows:

Table 1. IDs of the virtual objects the user wants to interacts with

	Glass	Bottle	Tap	Wheelchair	Closet door	Pot
Object ID	1	2	3	4	5	6

Within the hidden Markov model five states were considered, each one corresponding to a different intention the user may have. These states are described below:

- S1 = “I want to pour water from bottle into glass”;
- S2 = “I want to drink water from the glass”;
- S3 = “I want to wash the glass”;
- S4 = “I want to open the upper closet”;
- S5 = “I want to poor water from glass to pot”.

The transition matrix contains the probability of system transitions from one state to another. The values in this matrix influence the correct evolution of the statistical system. For example in Tab. 2, the system can easily evolve from S1 to S2, S2 to S3, S3 to S4, and S4 to S5, but it is unlikely that it will evolve from S5 to S1, S2 or S3, since the probabilities of these transitions were set to 0.1, in order to prevent the system from evolving in those states.

Table 2. System transition matrix of the hidden Markov model considered

	S1	S2	S3	S4	S5
S1	0.5	0.2	0.05	0.05	0.2
S2	0.3	0.4	0.2	0.05	0.05
S3	0.2	0.05	0.4	0.3	0.05
S4	0.1	0.1	0.1	0.3	0.4
S5	0.1	0.1	0.1	0.4	0.3

The sequence of input data is compared with the values of the observations/emission matrix, which has a unique set of data for every state. Thus, if a sequence of input data matches the emission matrix of a specific state, the probability that the system might evolve in that state is evaluated, and if it is high enough, the system will evolve.

The same data from Tab. 2 and 3 are reproduced in a more intuitive representation in Fig. 4 and 5. If counted, a total of 13 transitions have real chances to take place. The other transitions represented with thin lines are common situations in which the system cannot go, either because it is not possible or it is not desirable.

Table 3. Observations/emissions matrix of each state of the hidden Markov model considered

	O(t)	O(t+1)
S1 (O1)	ID1	ID2
S2 (O2)	ID1	ID4
S3 (O3)	ID1	ID3
S4 (O4)	ID4	ID5
S5 (O5)	ID1	ID6

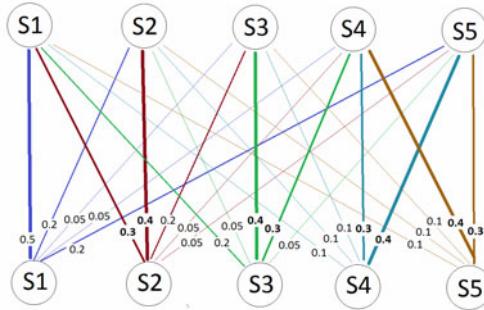


Fig. 4. Visual representation of the chosen states transitions probabilities of the hidden Markov model considered. Thicker lines and the associated values indicate higher probabilities of state transition. The representation is unidirectional, upwards.

In Fig. 5, each observation/emission has a highest correspondence connection with one state, which means it is most probably that the specific emission is produced by that particular state. These emissions are compared with the sequences of incoming data.

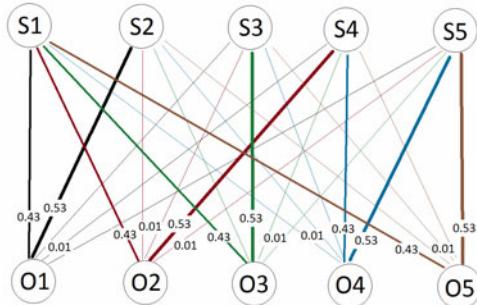


Fig. 5. Visual representation of the correspondence between observations/emissions and system states. Each observation has a most significant correspondent state.

Given a particular sequence $\text{seq}=[1,2,1,4,1,3,4,5,1,6]$, the system will evolve as expected from one state to the other. The `estimatedStates` vector demonstrates the reaction of the system to a particular user input sequence:

```
>> seq=[1,2,1,4,1,3,4,5,1,6]
seq =
1 2 1 4 1 3 4 5 1
estimatedStates=hmmviterbi(seq,trans,emis)
>> estimatedStates =
1 1 1 2 2 3 4 4 5
```

This example corresponds to the case in which the users gaze towards the glass, than towards the bottle, glass, wheelchair, glass, tap, wheelchair, closet door, glass and eventually pot. The estimated states show that when the users gaze sequentially

over the glass and bottle, the attentive interface determines that they want to pour some water from bottle to glass, which is correct. As it can be seen, detection of intentions in this way is simple and precise.

To go further with the analysis of system's response, given a sequence of incoming data which match two not directly transmittable states, it is noticeable that the system goes in an intermediate state until it reaches the final matched state:

```
>> seq=[1,2,1,3]
seq =
1 2 1 3
estimatedStates=hmmviterbi(seq,trans,emis)
>> estimatedStates =
1 1 2 3
```

If we consider the description of states S1, S2 and S3 involved in this example, it is a matter of logic that is not a common procedure to wash the glass if it was clean and some water was just poured into it. It is more likely that first the user should drink that water, and afterwards it might be necessary to have it cleaned. If direct transition is not possible, the system can evolve in intermediate states from which it can go further. As the number of states implemented is higher, more appropriate actions can be triggered in case of such direct transitions, which are either not recommended either impossible to be performed. Thus, the attentive interface will be able not only to understand what the user wants, but also to suggest alternatives when certain tasks are difficult to complete. To ensure a more accurate response, the system can wait for a longer sequence of inputs from the user (five user inputs, or more). In this way, if erroneous selections are made, the system can reject them very efficiently and infer the correct user intention.

5 Design of Experiments

The goal of the conducted experiments was to asses the effectiveness of the attentive user interface implementation based on hidden Markov model. The final purpose was to evaluate the accuracy of intention detection algorithm and also whether this interface is more natural compared to a command based interaction interface. 10 subjects have tested the two interfaces mentioned and at the end each one filled in a questionnaire about their experience of interacting with the virtual objects. Half of them were informed about the possible operation they could perform with virtual objects through the attentive interface. A description of all the five possible states was presented to these subjects, so they became aware about the possibilities and limitations of this interface. The other subjects were left to discover these features all by themselves during the experiments. They were all informed about the functionality of the command based interaction interface. This interface features instant selection of virtual objects, if the user gazes towards them. At the end they all answered a common set of questions, and in addition the five subjects who were not informed about the possibilities of the attentive interface, were asked whether their first impression of discovering it's features by themselves was positive or negative. The common requirement for all users were: 1) to count for the overall number of failed intention detections; 2) to compare which interface is more natural and less obtrusive.

6 Discussion of Results

The task of the users was to perform selections of virtual objects with the two interaction interfaces developed. After repeating the tests three times, those five subjects who were previously informed about the features of the attentive interface, managed to reach an 88% success rate, with an average of 1.8 failed intention detections per subject. The numbers are based on the answers they filled in the questionnaire after they completed the tests. The selection procedures made through the command based user interface reached a 100% precision, mostly due to the sufficiently distant arrangement of objects in the virtual environment. The answers to the second question indicates that the attentive interface based on the hidden Markov model implementation reacts more naturally and does not divert their attention from the surrounding environment. Some interesting user's comments state that it is fun to connect two objects with the eyes and see the system detecting what they can do with them. Those five users not aware about the features of the attentive interface reported a higher rate of failure, as some of them were expecting different possible operations between the gazed objects than the ones implemented in the system. Thus, the average failed detections were 5.6, with a corresponding 62% success rate. Some of them were not sure about the exact number of failed detections, as they were diverted from counting by the surprising discovery of unexpected connections between different objects. The answers to the second question state that for some users, the attentive interface is more natural, while for some others the difference is insignificant. They all confirmed that the attentive interaction interface is more intuitive than the command based interface.

Although this current implementation of an attentive user interface based on hidden Markov model is a rough one, the users of the system still managed to score an amazing success rate of 88%. This is encouraging as the possibilities of improving the accuracy are numerous.

Acknowledgment. This work was supported by the Romanian National University Research Council (CNCSIS-UEFISCDI), under the Grant INCOGNITO: Cognitive interaction between human and virtual environment for engineering applications, Exploratory Research Project PNII – IDEI 608/2008.

References

1. Majaranta, P., Räihä, K.J.: Twenty years of eye typing: systems and design issues. In: Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA 2002), pp. 15–22. ACM Press, New York (2002)
2. Selker, T.: Visual attentive interfaces. *BT. Technol. J.* 22(4), 146–150 (2004)
3. Prendinger, H., Ma, C., Yingzi, J., Nakasone, A., Ishizuka, M.: Understanding the effect of life-like interface agents through eye users' eye movements. In: Proceedings of Seventh International Conference on Multimodal Interfaces (ICMI 2005), pp. 108–115. ACM Press, New York (2005)
4. Eichner, T., Prendinger, H., André, E., Ishizuka, M.: Attentive presentation agents. In: Pelachaud, C., Martin, J.-C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) IVA 2007. LNCS (LNAI), vol. 4722, pp. 283–295. Springer, Heidelberg (2007)

5. Vertegaal, R.: Designing attentive interfaces. In: Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA 2002), pp. 22–30. ACM Press, New York (2002)
6. Prendinger, H., Hyrskykari, A., et al.: Attentive interfaces for users with disabilities: eye gaze for intention and uncertainty estimation. *Univ. Access Inf. Soc.* 8, 339–354 (2009)
7. Hyrskykari, A., Majaranta, P., Räihä, K.-J.: From gaze control to attentive interfaces. In: Proceedings of HCII 2005. Erlbaum, Mahwah (2005)
8. Shimojo, S., Simion, C., Shimojo, E., Scheier, C.: Gaze bias both reflects and influences preference. *J. Nat. Neurosci.* 6(12), 1317–1322 (2003)
9. Qvarfordt, P., Zhai, S.: Conversing with the user based on eyegaze patterns. In: Proceedings of the ACM CHI 2005 Conference on Human factors in Computing Systems, pp. 221–230. ACM Press, New York (2005)
10. Bailly, G., Raidt, S., Elisei, F.: Gaze, conversational agents and face-to-face communication. *J. Speech Communication* 52, 598–612 (2010)
11. Selker, T., Burleson, W., Scott, J., Li, M.: Eye-Bed. In: Workshop on Multimodal Resources and Evaluation in conjunction with the Third International Conference on Language Resources and Evaluation, LREC (2002)
12. Toet, A.: Gaze directed displays as an enabling technology for attention aware systems. *J. Comp. in Human. Beh.* 22, 615–647 (2006)
13. Jacob, R.J.K.: 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: Empowering People, New York, pp. 11–18 (1990)
14. Applied Sciens Laboratories, <http://www.asleyetracking.com>
15. 3d models on turbosquid, <http://www.turbosquid.com>
16. Barbuceanu, F., et al.: Evaluation of the average selection speed ratio between an eye tracking and a head tracking interaction interface. In: 2nd Doctoral Conference on Computing Electrical and Industrial Systems, Costa de Caparica, Portugal, pp. 181–186 (2011)
17. Cappe, O., Moulines, E., Ryden, T.: Inference in Hidden Markov Models. Springer, Heidelberg (2009)

A Low-Cost Natural User Interaction Based on a Camera Hand-Gestures Recognizer

Mohamed-Ikbel Boulabiar¹, Thomas Burger², Franck Poirier³, and Gilles Coppin¹

¹ LAB-STICC, Telecom-Bretagne, France

boulabiar@gmail.com, gilles.coppin@telecom-bretagne.eu

² LAB-STICC, University of Bretagne-Sud, France

³ VALORIA, University of Bretagne-Sud, France

{thomas.burger, franck.poirier}@univ-ubs.fr

Abstract. The search for new simplified interaction techniques is mainly motivated by the improvements of the communication with interactive devices. In this paper, we present an interactive TVs module capable of recognizing human gestures through the PS3Eye low-cost camera. We recognize gestures by the tracking of human skin blobs and analyzing the corresponding movements. It provides means to control a TV in an ubiquitous computing environment. We also present a new free gestures icons library created to allow easy representation and diagramming.

Keywords: natural gesture interaction, low-cost gesture recognition, interactive TV broadcast, ubiquitous computing.

1 Introduction

HCI research focuses more attention than before on enhancing the user experience regarding human-display interaction. This area of research focuses on making interaction more natural, so that it is not necessary to click on buttons and to touch screens. To do so, automatic gesture recognition is an interesting paradigm [3], [13], [6]. However, in everyday life, for interaction with home devices having embedded computational power, such as interactive TVs, one does not make benefit from these new interaction paradigms yet.

We present in this paper an approach that allows the detection and the interpretation of gestures of a TV spectator from a simple and low cost camera. This work takes place in the context of the French FUI¹ Project RevTV, the aim of which is to add new interaction techniques, so that telespectators take actively part to the TV broadcast. The final objective of this project is to control the animations of an avatar that should be inserted within a TV show beside the presenter or actor. By now, the major scenario that we rely on, corresponds to some educational game where a pupil controls her or his avatar which interacts in a broadcast program led by a “real” TV animator. During such kind of scenarios, commands interaction like pointing, selecting and moving, as well as natural gesture animation, are required.

¹ Fond Unique Interministeriel.

The paper is organized as following. Section 2 presents the study held on gestures semantics, sources, and taxonomies, and includes the presentation of the free icons library created and used. Section 3 shows the technical details of handling the camera input for gestures recognition. Section 4 shows the modes where the gestural information generated are used with some screenshots of the running application. Section 5 answers the usability and natureness question of the use of gestures specially in an ubiquitous environment. Finally section 6 discusses the future of the work and the possible integration of other multimodal modules.

2 Gesture Semantics

Symbolic gestures, such as emblems, play an important role in human communication as they can fully take the place of words. These gestures are processed by the same area of the human brain as the spoken language [17]. Hence, these gestures do not come alone, and they are often combined with speech, or with cognitive activities, as they are classically used to add more precision on the details of a talk.

We have extracted some semantics of such type of gestures, by the analysis of the movements of a user explaining a story, and by searching for a relation between gestures and the part of the story being told [9]. After extraction, most of gestures semantics can be divided into two parts for their possible future use as:

Animation Gestures used to animate a virtual avatar.

Command Gestures used to launch a specific predefined action.

2.1 Gestures Taxonomies

A rapid look the possible gestures human can perform [16], lead to the conclusion that the amount of possibilities is tremendous. Nonetheless, the gestures can be classified according to various criteria not only to simplify their recognition, but also to allow their reproducing on an avatar to putting tags on them for future analysis [10]. Here, we consider a taxonomy based on the context of the gestures as well as on their types:

According to McNail [9], it is possible and useful to distinguish between these gestures. Here follows a possible taxonomy:

Gesticulation is a motion that embodies a meaning reliable to the accompanying speech. It is made chiefly with the arms and hands but is not restricted to these body parts.

Speech-linked gestures are parts of sentences themselves, the gesture completes the sentence structure.

Emblems are conventionalized signs, such as thumbs-up or the ring (first finger and thumb tips touching, other fingers extended) for “OK.”

Pantomime is dumb show, a gesture or sequence of gestures conveying a narrative line, with a story to tell, produced without speech.

In our work, we focus on Emblems for Command Gestures, and on Gesticulations for Animation Gestures.

2.2 Gestures Library

In order to identify relevant gestures we have analyzed some talk-shows, and we have extracted the most frequent gestures being performed. According to our scenario of use in the context of interactive TV (RevTV project), we need some gesture to trigger commands, as well as other gestures to implement natural interaction with the avatar. We have created a vector based gestures icons to represent them. These 4 icons, represented in figure 1, are part of the **ishara** library [1]. They represent the gestures supported by the software of our application. We provide a movement mode, a multitouch mode, a command mode, a hand fingers recognition, and, finally, a cursor movement mode (see Section 5 for detailed explanation of these modes).



Fig. 1. Family of supported gestures represented using the free gestures icons library from [1]

These five basic gestures modes are already supported in our system as a first step. They will be extended depending on the context fulfilling the requirements of our scenario.

3 Technical Work

Figure 2 is a description of the pipeline of the modules used to extract gestural information. Each part is described in details in the next subsections.

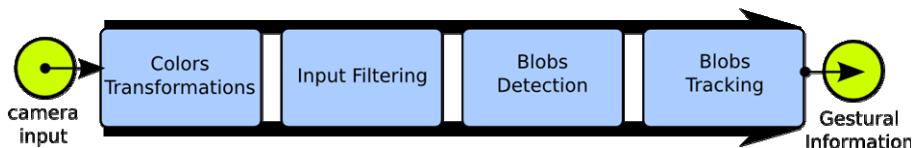


Fig. 2. Pipeline of the work components showing the routing and transformation of the camera information until getting the gestural information ready to use in the scenario

3.1 Camera Input

We have chosen a low-cost camera that can be integrated with other components in the future. The PS3Eye camera matched our expectations with a 640x480 resolution combined with a 60 frames-per-second capability. This camera costs about 40\$, which is a reasonable price for a potential market deployment.

3.2 Colors Transformations

The video processing which depend on the OpenCV library [4] is based on several modules, such as illustrated in figure 2. First, a basic but adaptive skin color segmentation of a large interval of human skin varieties is performed [2]. We also perform histogram equalization technique to better enlarge the colors to all the space available. Then we transform the color space from BGR, which come from the camera to YCrCb, as it is the most adapted one to skin segmentation. Other color spaces like HSV or CIE-Lab can be used for skin color segmentation, but the transformation from BGR either takes more time (the transformation we use is linear) or needs more parameters in the next step [2].

Space	BGR	HSV	YCrCb	CIE Lab
Param.	R>95, G>40, B>20, Max{R,G,B}-Min{R,G,B}<15 abs(R-G)>15, R>G, R>B	H[0.4 .. 0.7] S[0.15 .. 0.75] V[0.35 .. 0.95]	Cb[77 .. 127] Cr[133 .. 173]	C[0 .. 65] I[0 .. 14]

3.3 Input Filtering

During this step, a 5x5 opening morphological kernel [15] is applied to smooth the results. Practically, it cleans the image from isolated pixels and holes inside skin zones.

3.4 Blobs Detection

Each blob is labeled in linear time using the contour tracing technique with the linear-time component-labeling algorithm [5], which requires 1 to 4 passes to recognize all blobs with their bounding contour. We identify the hands and the head by a calibration focused on their first position in the screen. We identify them and add a tag to their respective blobs.



Fig. 3. A representation of different human skin color variations

3.5 Blobs Tracking

After having the hands and head blobs identified, an efficient tracking method based on appearance model is used [14]. Let us note that it handles occlusion problems. We have used the cvblob implementation [8] for that algorithm which stores the measures of bounding boxes distance and provides tracks instead of just blobs. Once the matrix of distance is defined, the following loops are executed to handle blobs changes:

- A loop to detect inactive tracks: those tracks with no blobs near.
- A loop to detect and create new tracks: those blobs without a track near.
- A loop to assign blobs to tracks. It makes clusters with blobs that are close and assign them to a track. In this step some tracks could merge in one.
- A last loop which check all inactive tracks to delete the old ones.

To better handle blobs, the bounding path of the blobs is simplified to a set of few points when needed. This is used to identify picks for the counter mode (see Section 5). Finally, a phase of calibration is triggered when the logic of tracking is lost or damaged.

4 Gestural Information and Recognition Modes

The output of the video processing is the input of the recognition module, which is based on the track location on the screen. Four different modes are defined according to the hand locations. The switch between them is based on the following grammar:

The movement mode: it is activated when two hands are close to each other, so that the user selects between 4 directions. This mode can be used as input for games to select between the directions.

The multi-touch mode: where we consider the similarities between 2D gestures in a tactile multi-touch context and gestures in the 3D space. To do so, we consider the hands blobs in a manner similar to that of two finger tips in the input of a multi-touch device. Based on that, we can recognize well known multi-touch gestures, such as Drag, Pinch/Zoom and Rotate with a consideration of a interaction centroid [7] but applied in our case.

The counter mode: counts the number of fingers in a preselected hand (the left and right hand are automatically discriminated) by counting the peaks in the simplified contour of the blob. This mode can again be used in games for kids.

The mouse mode: it is used to control a cursor to select something in an arbitrary place on the screen. With this mode, we only get the input from a sub zone in the screen and map it to the full screen. We use it in a similar way to a computer touchpad. The user can validate clicks using a tempo on another zone.

5 Usability and Naturalness in an Ubiquitous Environment

The usability of our system should be compared to other systems in an ubiquitous environment. In our case, we have a TV instead of a PC. The gestures in such environment, and specially for children game scenario, are made for a short period of time, and doesn't require a good usability, even if we can recognize gestures in real-time.

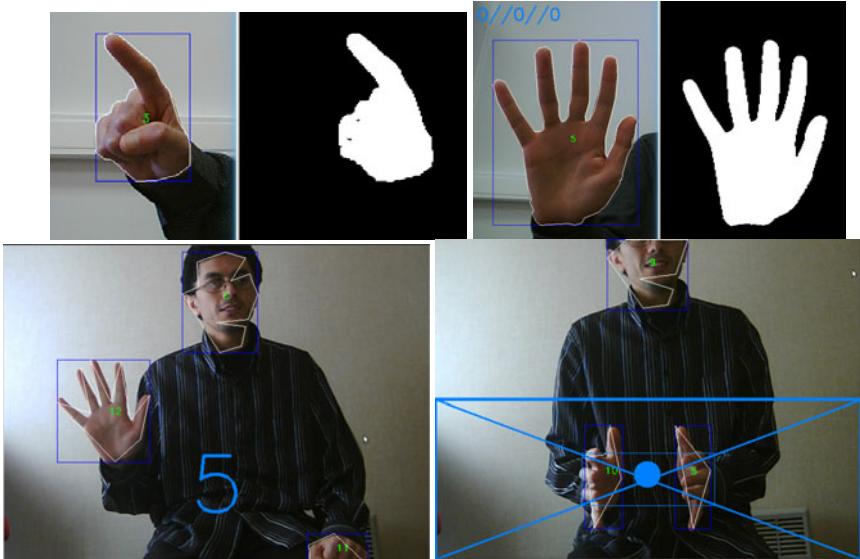


Fig. 4. A sub-list of recognition modes showing pointing, number of hand fingers counting and the movement mode

The Naturalness of gestures is supported by the analysis of those to be supported and the preference of gesticulations. Gesticulations are the most common gestures, so they inherit from this their ability to be produced with less complication.

Emblems gestures which englobe commands are chosen to be produced in more difficult positions to allow easy discrimination from gesticulations. Even with these extractions, the naturalness aspect of the gestural interaction itself is still objected. [11]

Spatial gestural interactions always lack from the feedback for the interaction. This affects also the usability because the user can no more be guided in space as he was with surface movement or simple physical joysticks.

6 System Integration of This Interaction and Future Work

Human brain Broca's area, identified as the core of the brain's language system, are not in fact committed to language processing, but may function as a modality-independent semiotic system that plays a broader role in human communication, linking meaning with symbols whether these are words, gestures, images, sounds, or objects [17]. This natural handling of multimodal communications gives an argument to build recognizer taking part of most of these modalities specially to remove ambiguous gesture meaning decision cases.

Our system only supports gestural interaction at the moment but the integration of other multimodal means are possible to fine tune the input and give the user better feedback. These possible evolutions of the system are possible:

Facial Recognition: our system does not take care of the facial expressions recognitions. A future support in this area can be added for more reactivity in pupil's games.

Voice Recognition: To speed up commands handling, we can use short voice keywords either to move from one mode to another, or to select object.

Haptic feedback: A special wearable vest can be used to allow more reactivity with the user. But this area is still lacking innovation because the haptic feedback is by area and not continue.

7 Conclusion

In this paper we have shown a proof-of-concept mechanism to recognize hand gestures then use them to control a TV or to take actions in a Game scenario for pupils. We have also provided a free library which can be used to represent gestures in other scenarios and other studies.

The evolution of our system could be possible in the direction of a multimodal environment in the case where we can be aware of the limits and myths [12]. But this evolution is a natural choice as new devices are getting more computational power and the ubiquitous environment is becoming a reality.

References

1. Ishara vector based and open gestures icons (2011),
<https://github.com/boulabiar/ishara>
2. Askar, S., Kondratyuk, Y., Elazouzi, K., Kauff, P., Schreer, O.: Vision-based skin-colour segmentation of moving hands for real-time applications. In: 1st European Conference on Visual Media Production (CVMF), March 2004, pp. 79–85 (2004)
3. Bolt, R.A.: Put-that-there: Voice and gesture at the graphics interface. In: Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1980, pp. 262–270. ACM Press, New York (1980)
4. Bradski, G.: The OpenCV Library. Dr. Dobb's Journal of Software Tools (2000)
5. Chang, F., Chen, C.J., Lu, C.J.: A linear-time component-labeling algorithm using contour tracing technique. Comput. Vis. Image Underst. 93(2), 206–220 (2004)
6. Derpanis, K.: A review of vision-based hand gestures. Tech. rep (2004)
7. Gorg, M.T., Cebulla, M., Garzon, S.R.: A framework for abstract representation and recognition of gestures in multi-touch applications. In: ACHI 2010: Proceedings of the 2010 Third International Conference on Advances in Computer-Human Interactions, pp. 143–147. IEEE Computer Society Press, Washington, DC (2010)
8. Linan, C.C.: cvblob, <http://cvblob.googlecode.com>
9. McNeill, D.: Gesture and Thought. University of Chicago Press (2005)
10. Neff, M., Kipp, M., Albrecht, I., Seidel, H.P.: Gesture modeling and animation based on a probabilistic re-creation of speaker style. ACM Trans. Graph. 27(1), 1–24 (2008)
11. Norman, D.A.: The way i see it: Natural user interfaces are not natural. Interactions 17(3), 6–10 (2010)
12. Oviatt, S.: Ten myths of multimodal interaction (1999)
13. Pavlovic, V.I., Sharma, R., Huang, T.S.: Visual interpretation of hand gestures for human-computer interaction: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence 19, 677–695 (1997)

14. Senior, A., Hampapur, A., Tian, Y.L., Brown, L., Pankanti, S., Bolle, R.: Appearance models for occlusion handling. *Image and Vision Computing* 24(11), 1233–1243 (2006)
15. Soille, P.: *Morphological Image Analysis: Principles and Applications*. Springer, Heidelberg (2004)
16. Streeck, J.: *Gesturecraft*. John Benjamins Publishing Company, Amsterdam (2009)
17. Xu, J., Gannon, P., Emmorey, K., Smith, J., Braun, A.: Symbolic gestures and spoken language are processed by a common neural system. *Proc. Natl. Acad. Sci. U.S.A.* 106(49), 20664–20669 (2009)

Head-Computer Interface: A Multimodal Approach to Navigate through Real and Virtual Worlds

Francesco Carrino^{1,2}, Julien Tscherrig¹, Elena Mugellini¹,
Omar Abou Khaled¹, and Rolf Ingold²

¹ College of Engineering and Architecture of Fribourg, Switzerland

{Francesco.Carrino, Elena.Mugellini, Omar.AbuKhaled}@Hefr.ch,

Julien.Tscherrig@Edu.Hefr.ch

² University of Fribourg, Switzerland

Rolf.Ingold@Unifr.ch

Abstract. This paper presents a novel approach for multimodal interaction which combines user mental activity (thoughts and emotions), user facial expressions and user head movements. In order to avoid problems related to computer vision (sensitivity to lighting changes, reliance on camera position, etc.), the proposed approach doesn't make use of optical techniques. Furthermore, in order to make human communication and control smooth, and avoid other environmental artifacts, the used information is non-verbal. The head's movements (rotations) are detected by a bi-axial gyroscope; the expressions and gaze are identified by electromyography and electrooculography; the emotions and the thoughts are monitored by electroencephalography. In order to validate the proposed approach we developed an application where the user can navigate through a virtual world using his head. We chose Google Street View as virtual world. The developed application was conceived for a further integration with a electric wheelchair in order to replace the virtual world with a real world. A first evaluation of the system is provided.

Keywords: Gesture recognition, Brain-Computer Interface, multimodality, navigation through real and virtual worlds, human-computer interaction, psychophysiological signals.

1 Introduction

Smart environments, augmented reality and human body used as controllers are just some examples of how the technology is getting introduced into our everyday life activities, increasing demand for natural ways of interaction and communication. On the one hand, we are observing a proliferation of interactive devices, user interfaces, and wearable technologies (such as Nintendo Wii®, Microsoft Kinect® and Play-Station Move®); on the other hand the human has to adapt, to learn, to master and to use those new devices.

In order to make human-computer interfaces truly natural, we need to develop technologies that track human movement, body behavior and facial expressions, and interpret this information in an “affective” way (this means taking into account also the

subject's emotional state)[1]. The gesture-based interaction is a natural way to interact and can represent a substitution/complement of other forms of communications or in special context (such as concerning impaired people interaction capabilities) [2].

The usual approaches to address the gesture recognition challenges go into the direction of computer vision and image processing [3-5]. These methods are limited by some typical environmental constraints such as sensitivity to lighting changes, reliance on camera position, etc. Moreover, the images elaboration is very expensive in terms of computer processing power, making the real-time analysis a difficult challenge [6].

Additionally psycho-physiological sensors such as Electroencephalogram (EEG), Electromyogram (EMG), Blood Volume Pressure (BVP) or Galvanic Skin Response (GSR) (just to mention some of the well-known technologies) give important information about the cognitive, affective or subject's health conditions.

Furthermore, non-verbal information, such as facial expression, gaze and gesture, plays an important role in human communication [7]. The exchange of non-verbal information is important in all forms of communication and, in some specific contexts (e.g. impaired people), it is even more important than verbal information.

According to this, our paper focuses on the Human-Machine-Interaction with particular attention to the non-optical, non-verbal techniques. In this work we interlace the concepts of Gesture Recognition, Context Awareness and Multimodal Interaction in order to enhance our communication and control capabilities, and offer impaired people new ways of interaction.

In order to validate our idea, this paper presents a prototype, called Virtual Move, which allows users to navigate through Google Street View using the head only. The herein adopted concept of "head" includes movements and expressions, as well as thoughts and emotions.

This paper is organized as follows: Section 2 describes the state of the art of related researches. Section 3 describes the presented prototype focusing on the conception, the adopted interface and the evaluation. Finally, section 4 concludes the paper and discusses future work.

2 Related Projects

Multimodal approaches are often adopted in gesture recognition systems. For instance face and gestures recognition are combined in [8], where the author describes three examples of systems developed using a design-for-all strategy: a gesture sensing devices to replace the mouse; a speech to lip conversion for the hearing impaired; a brain-computer interaction based on the analysis of electroencephalographic recordings in response to gestural movements for severely disabled people.

Hands-free manipulation systems have also been studied in several works. Takahashi [7] presents a non-verbal approach based on bio-potential signals (electrooculography (EOG), electromyography (EMG)) for simple gesture recognition and he proposes a prototype that aims to control a moving object in a virtual space. A similar approach is adopted in [9], where the two produced experiments aim to control at first a walkthrough application in a 3D virtual space and then an active camera. In [10] Hashimoto uses expressions and gaze to control an electric wheelchair in a non-verbal oriented approach.

A combination of mechanical (accelerometers) and physiological (EMG) sensors is used [11, 12] to identify hand gestures in order to control various devices or in substitution of videogames controllers.

Several works [1, 5, 13] propose head gestures (e.g. nodding, shaking, tilting, etc.) recognition using a computer vision approach or speech recognition techniques.

With respect to these works we can state that multimodal use of different sensing technologies can allow an effective and functional interaction for people with limited mobility/coordination abilities, or under specific conditions where it is not possible to use neither the whole body nor the voice.

However, what is currently missing is a system that jointly exploits all the interaction capabilities offered by the head in order to allow users to interact using non-optical and non-verbal techniques. Especially brain signals, such as the EEG, are often not considered.

This paper presents a system that allows the user to communicate and control the navigation in a virtual environment using head movements, facial expressions, thoughts and emotional states. The proposed approach doesn't make use of optical or verbal techniques.

3 Virtual Move Concept and Prototype

In this paper we describe our prototype, called Virtual Move, a multimodal application that aims to allow the users to navigate in Google Street View (from now GSV) using the head. EEG, EMG, EOG and a biaxial gyroscope signals are collected using the Epoch Emotiv headset [14] in order to exploit all the communication capabilities of the human head (i.e. expressions, head's movements, emotions and thoughts that, from now on, we will refer to as "head's actions"). Next sections show details about the conception and the interface that allow the user to communicate in a multimodal way with the application. Finally an evaluation about the used technology and the developed application is provided.

3.1 Concept

Virtual Move is based on the following main ideas:

- Multimodality – in order to exploit all the human head interaction capabilities (thoughts and emotions included).
- Flexibility – in order to allow a large number of users to use the application, Virtual Move has to adapt to the different possible users' needs. In the case of a subject that can not perform a specific movement or expression, the application should be able to replace the head's action with another, or add redundancy allowing the user to use at the same time different head's actions in order to transmit the same command to the navigator.
- Reusability (from the software engineering point of view) – that facilitate the forthcoming adaptation to an electric wheelchair, in order to shift from a virtual world to the real world.

In order to do that, Virtual Move is placed as bridge between the EmoEngine that analyzes the data coming from the headset, and the GSV API (see Fig. 1). Moreover, Virtual Move functions as an interface including on its Main View the GSV map's navigator and other tools to make the navigation using the head easier.

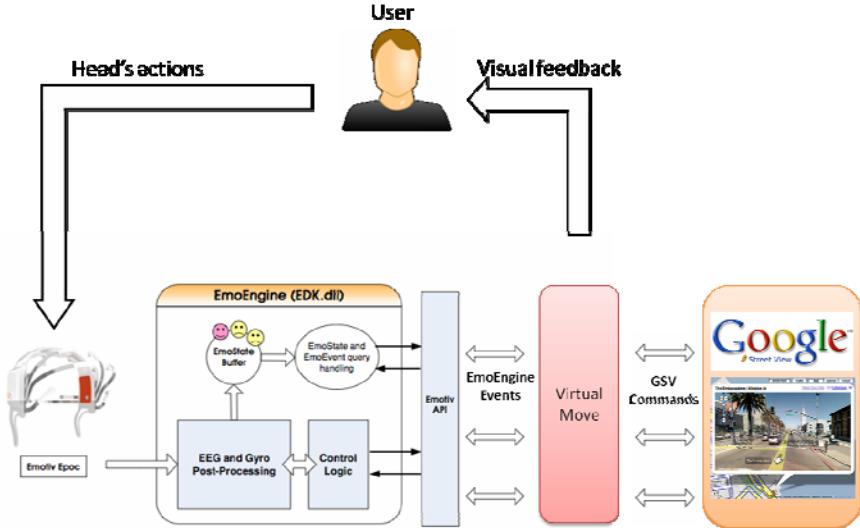


Fig. 1. The Communication Flows between the user, the Epoch headset, EmoEngine, Virtual Move and Google Street View

The EmoEngine is the main application provided with the Emotiv Epoch headset. This software does the whole signal processing and then creates “events”. These events are related to all the needed information: from the detected expressions to the gyroscope information, from the detected emotion to the “thoughts”. Obviously the EmoEngine can not recognize each kind of thought. The detected cognitive events are related to the so called Motor Imagery [15-17]. The Emotiv application allows the user to train the system to recognize maximum four gestures simultaneously in addition to the neutral state (no movements, no thoughts). The performance of the classifier and then the usability of this interaction modality are strongly related to the entrainment rate achieved by the user. However, according to the tests we have performed, using more than two cognitive actions is practically impossible, due to the low precision of the EEG sensors.

The events created by the EmoEngine are read by Virtual Move and sent to the GSV API allowing the user to navigate as he wishes.

Fig. 1 shows the communication flows between Virtual Move, EmoEngine and Google Street View. The Emotiv Epoch headset recognizes the head's actions processing data such as gyroscope coordinates and EEG, EMG, EOG signals. From the biaxial gyroscope it can detect heads movements such as nodding and shaking (tilting is not recognized); EEG gives information about emotional state and cognitive tasks; EMG detects muscular face activity giving information about the user's expressions,

while the EOG detects the eyes activities. It is important to remark that all the psychophysiological signals (EEG, EMG and EOG) are detected by the same type of sensors. EMG and EOG signals are much stronger than EEG signals (i.e. measured EMG/EOG potentials range from about 100 μ V to 100 mV, while the EEG potentials range between 0.5-100 μ V). This means that using commercial EEG devices, cognitive signals are less accurate. For this reason, applications that make use of EEG data have to be more error-tolerant. In our application it is suggested to use cognitive actions and emotions carefully or for less important activities so as not to impair navigation. For the same reason, in a pure BCI approach, EMG and EOG are considered as annoying noise.

After the signal processing the EmoEngine generates events related at the detected head's actions. These events are handled by Virtual Move, which transforms them into commands for GSV. For example, a left wink can be associated to a GO_LEFT command, and a right wink to a GO_RIGHT. Head movements and cognitive actions are handled in the same way. The only difference concerns emotions that trigger GSV commands when they exceed a given threshold. Based on the same principle, we can easily imagine replacing the GSV API with the API of an electric wheelchair. However, the associations and thresholds configurations are managed in the Virtual Move Settings Windows, independently from the type of device (GSV, electric wheelchair or any other physical or virtual device).

3.2 Interface

Virtual Move is composed by two principal modules: the Settings Windows (fig. 2) and the Navigation (or Main) Window (fig. 3).

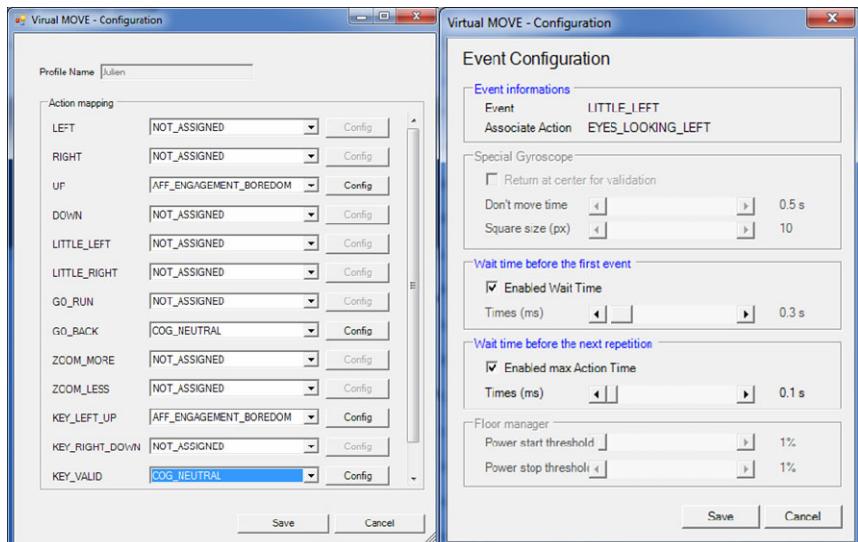


Fig. 2. Settings Windows

In our application each command made available from the Google Street View (Google Maps) API (i.e. "GO_FORWARD", "GO_BACK", "TURN_LEFT", "TURN_RIGHT", "UP", "DOWN", etc.) can be associated with the user's head's actions detected by the Emotiv headset. Each user can set his profile, however a default profile is provided. These associations are set in the Settings Windows.

In these panels is embedded the application flexibility and multimodality. In fact, each Google Street View command can be associated with one or more head's actions going in the direction of a "Concurrent" multimodal approach (CASE model [18]). This was done because patients with affected motor functions may have difficulties with some kind of actions. For instance, in the case of a user unable to perform a rotation to the left, the application allows him to use another movement (e.g. nodding) as well as another modality (e.g. a facial expression). Moreover, modulating thoughts or emotions in order to continuously command a system requires constant concentration. Hence, this could compromise the reliability of the cognitive approach. Combining inputs from several modalities adds redundancy to the system (CARE model [19]) allowing user to interact easier.

Moreover, in the Settings Windows, the user can set the thresholds for commands triggered by the emotions. For example, a scroll bar allows choosing the necessary relax level to produce an action such as make appear on the screen a virtual keyboard to type (always using head's actions) the name of the city to visit.

In Virtual Move the use of the gyroscope can become pretty advanced. In fact, the user can chose to trigger a GSV command not only by a single head movement but also by a sequence of movements (e.g. a possible sequence can be "up-down-up-down"). This approach considerably increases the interaction possibilities given from the gyroscope allowing the association of several GSV commands.

Other parameters can also be modified for a refined configuration.

The Navigation Window is where the user, principally, can see the Google Street View maps and interact with them by his head's actions. In this window besides the maps, all the feedbacks from the selected modalities are showed to the user. On the left (see fig. 3) several panels, bars and graphics allow the control of the gyroscope position, the emotion level and, in general, the detected head's actions.



Fig. 3. User interacting with the Navigation Window

In this window it is also possible to enable a virtual keyboard, making possible to type the name of the city and/or street to visit.

Virtual Move offers to the user two interaction possibilities: gyro-A and gyro-B. The first one is strongly related to visual feedbacks. The user, watching in a rectangular zone, can see in real time the gyroscope coordinates (visualized as a red dot) and act consequently. Touching with the red dot different regions, separately or in sequence, he can trigger the desired GSV command. This approach presents three main problems: the user is forced to focus on the gyroscope panel; the user's head is never free to move; the drift present in the gyroscope signals (a typical problem of inertial systems). For this reason we developed the second interaction possibility. In gyro-B modality, instead of the gyroscope position we consider the velocity. The command is detected only when the head is moving faster than a defined threshold (configurable in the Settings Windows). This approach avoids the previous problems but we have to consider that people with coordination difficulties could not be able to perform such kind of movements. For this reason both approaches are available. However, tests on healthy people have demonstrated a preference for the second approach.

3.3 Headset and Prototype Evaluation

During the prototype development we did two evaluations. Each one was composed by a first part where the users were testing the system following a specific test protocol, followed by a questionnaire. In the first evaluation, we tested the adopted technology, the Emotiv Epoch headset, with thirteen healthy subjects. In the second evaluation we tested the Virtual Move application on a smaller sample. For the sake of brevity we will show only the result of the second, since testing the device is not the focus of our work. However, it is important to highlight some points. Testing technologies such as the Emotiv Headset poses several difficulties. Firstly, we are dealing with the brain whose behavior is quite variable from person to person [20]. Secondly, we don't know what happens within the EmoEngine "black box", because we do not have information about the signal processing actuated by EmoEngine. For all these reasons the results of these experiments were not unequivocal. What we can affirm is, on the one hand, a concrete difficulty to reproduce/detect a cognitive action without a long training for the user. On the other hand, the possibility to interact using emotions and thoughts is very appealing and all the subjects marked the system as "not frightening".

Virtual Move application was tested with five healthy subjects. The starting configuration was the same for all the users: the same head's actions were related to the same GSV commands. After five minutes of free navigation we switched the gyroscope to the second modality (gyro-B, see the previous section). After five minutes we asked the subjects to adapt the settings according to their preferences. Finally, we asked them to attain three specific locations.

Results came as the following:

- All the subjects attained the requested locations.
- The most relevant commands ("GO_FORWARD", "GO_BACK", "TURN_LEFT" and "TURN_RIGHT") were assigned to the modality felt more reliable: the gyroscope. Followed by expressions, emotions and cognitive actions.

- Four subjects out of five preferred the gyro-B modality. The fifth didn't care.
- Relax was the only emotion used (in order to make appear and disappear the virtual keyboard, as in the default configuration)
- Since the required relax level to trigger the command was often achieved after closing the eyes, it was suggested to add an acoustic feedback for this type of interaction.
- Cognitive interactions were used with skeptical curiosity.
- The virtual keyboard was described as "functional but boring to use".
- The subjects were doubtful about an application in the real world.
- Mouse and keyboard, if available, remain a better solution.

In conclusion, the Virtual Move tests were satisfying, showing good system usability and the application resulted flexible and user friendly. However, as expected, our subjects expressed preference for the traditional "mouse and keyboard" interaction, waiting for a strong improvement of the brain-related technology. This was not surprising, being our test's subjects healthy. An important future step will be testing the application with severely impaired people and compare the result.

4 Conclusion and Future Work

In this paper we proposed a novel multimodal approach to navigate within a virtual world using the head. In order to avoid problems related to computer vision, the proposed approach does not make use of optical techniques. Furthermore, in order to avoid other environmental artifacts and make human communication and control smoother the used information is non-verbal.

Finally a prototype was developed and evaluated, as well as the adopted technology.

As next step we would like to use the Virtual Move application to control an electric wheelchair. Obviously, errors and inaccuracies assume greater importance in the real world. In addition, this technology (overall the BCI aspect) must be firstly accepted and understood by the end user. In this sense our tests were encouraging, with the shared subjects' opinion that BCI technologies are not frightening. However, new tests involving impaired people will be very useful for the transition to the real world.

Finally, aiming to achieve more independence from the EmoEngine "black box", the next projects will interface raw data directly.

References

1. Gunes, H., Piccardi, M., Jan, T.: Face and body gesture recognition for a vision-based multimodal analyzer. In: Proceedings of the Pan-Sydney area, vol. 36, pp. 19–28 (2004)
2. Mitra, S., Acharya, T.: Gesture Recognition: A Survey. IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews) 37, 311–324 (2007)
3. Davis, J., Shah, M.: Visual gesture recognition. Vision, Image and Signal Processing. In: IEE Proceedings, IET , pp. 101–106 (2002)
4. Je, H., Kim, J., Kim, D.: Vision-Based Hand Gesture Recognition for Understanding Musical Time Pattern and Tempo. In: 33rd Annual Conference of the IECON 2007, pp. 2371–2376. IEEE Industrial Electronics Society (2007)

5. Ng, P.C., De Silva, L.C.: Head gestures recognition. In: International Conference on Image Processing, Proceedings, pp. 266–269. IEEE, Los Alamitos (2002)
6. Keir, P., Payne, J., Elgoyhen, J., Horner, M., Naef, M., Anderson, P.: Gesture-recognition with Non-referenced Tracking. *The Computer Journal*, 151–158 (2006)
7. Takahashi, K., Nakauke, T., Hashimoto, M.: Remarks on Simple Gesture Recognition of Bio-Potential Signals and Its Application to Hands-Free Manipulation System. In: TENCON 2006 - 2006 IEEE Region 10 Conference, pp. 1–4 (2006)
8. Reilly, R.B.: Applications of face and gesture recognition for human-computer interaction. In: Proceedings of the Sixth ACM International Conference on Multimedia Face/Gesture Recognition and their Applications - MULTIMEDIA 1998, pp. 20–27 (1998)
9. Takahashi, K., Hashimoto, M.: Remarks on EOG and EMG gesture recognition in hands-free manipulation system. In: 2008 IEEE International Conference on Systems, Man and Cybernetics, pp. 798–803 (2008)
10. Hashimoto, M., Takahashi, K., Shimada, M.: Wheelchair control using an EOG- and EMG-based gesture interface. In: IEEE/ASME International Conference on Advanced Intelligent Mechatronics, pp. 1212–1217 (2009)
11. Zhang, X., Chen, X., Wang, W.-h., Yang, J.-h., Lantz, V., Wang, K.-q.: Hand gesture recognition and virtual game control based on 3D accelerometer and EMG sensors. In: Proceedings of the 13th International Conference on Intelligent User Interfaces - IUI 2009, vol. 401 (2008)
12. Chen, X., Zhang, X., Zhao, Z.-Y., Yang, J.-H., Lantz, V., Wang, K.-Q.: Hand Gesture Recognition Research Based on Surface EMG Sensors and 2D-accelerometers. In: 11th IEEE International Symposium on Wearable Computers, pp. 1–4 (2007)
13. Morimoto, C., Yacoob, Y., Davis, L.: Recognition of head gestures using hidden Markov models. In: Proceedings of 13th International Conference on Pattern Recognition, pp. 461–465 (1996)
14. Emotiv - Brain Computer Interface Technology, <http://www.emotiv.com/>
15. Pfurtscheller, G., Neuper, C.: Motor imagery and direct brain-computer communication. *Proceedings of the IEEE* 89, 1123–1134 (2001)
16. Pfurtscheller, G., Neuper, C.: Motor imagery activates primary sensorimotor area in humans. *Neuroscience Letters* 239, 65–68 (1997)
17. Scherer, R., Lee, F., Schlögl, A., Leeb, R., Bischof, H., Pfurtscheller, G.: Towards self-paced (asynchronous) Brain-Computer Communication: Navigation through virtual worlds. *IEEE Transaction on Biomedical Engineering* 55, 675–682 (2008)
18. Dumas, B., Lalanne, D., Oviatt, S.: Multimodal interfaces: a survey of principles, models and frameworks. *Human Machine Interaction*, 3–26 (2009)
19. Nigay, L., Coutaz, J.: A design space for multimodal systems: concurrent processing and data fusion. In: Proceedings of the INTERACT 1993 and CHI 1993 Conference on Human Factors in Computing Systems, pp. 172–178. ACM Press, New York (1993)
20. Guger, C., Edlinger, G.: How many people can control a brain-computer interface (BCI). In: Proc. BrainPlay, pp. 29–32 (2007)

3D-Position Estimation for Hand Gesture Interface Using a Single Camera

Seung-Hwan Choi, Ji-Hyeong Han, and Jong-Hwan Kim

Department of Electrical Engineering, KAIST,
Gusung-Dong, Yusung-Gu, Daejeon, Republic of Korea
`{shchoi, jhhan, johkim}@rit.kaist.ac.kr`

Abstract. The hand gesture interface is the state of the art technology to provide the better human-computer interaction. This paper proposes two methods to estimate the 3D-position of the hand for hand gesture interface using a single camera. By using the methods in the office environment, it shows that the camera is not restricted to a fixed position in front of the user and can be placed at any position facing the user. Also, the reliability and usefulness of the proposed methods are demonstrated by applying them to the mouse gesture recognition software system.

Keywords: Position Estimation, Hand Gesture Interface, Human Computer Interaction.

1 Introduction

Recently, various hand gesture interfaces have been developed. Among them, the vision-based hand gesture interface with a webcam is efficient and popular because it only needs a webcam which is already very common. There are mainly two approaches in the vision-based hand gesture interface, i.e. 3D-model-based and 2D-appearance-based. 3D-model-based approach fits the hand image to a 3D hand model, which is already constructed, using extracting hand features or hand outlines from the camera image [1][2][3]. This approach estimates the hand pose highly accurately, but it takes a long processing time. 2D-appearance-based approach directly compares the input hand image to the database images [4][5]. This approach takes a short processing time, but it does not estimate the hand pose accurately when the wrist or the forearm is moving. The hand pose is easily coupled with information such as software commands like click or drawing. However, users have to study each hand pose corresponding to each command to use the software.

The hand position, on the other hand, is more intuitive to control the software than the hand pose. Additionally, most of software is also available to use the interface with the 3D hand position because the 3D-position can be easily translated to a 2D-position and additional state information on mouse commands. There are several methods to find out the 3D-position, such as by using a depth cam or a stereo cam, but they are still not popular among most of the users. This paper proposes two different methods to detect the 3D-position of a hand by using a single camera for the hand

gesture interface. The proposed methods use homography and the neural network. The camera is not restricted to a fixed position in front of the user and can be placed at any position facing the user.

This paper is organized as follows: Section 2 describes the hand detection and proposed methods for estimating 3D-position of the detected hand. In Sections 3, the experimental results are described. Finally, concluding remarks follow in Section 4.

2 3D-Position Estimation

The proposed method to estimate 3D-position of the hand consists of two processes. The first one is hand detection in a camera image. The camera image contains information about the hand such as its 2D-position, direction and pose. Among them, the 2D-position and the size of the hand are used. The other one is the 3D-position estimation of the detected hand. This paper proposes two methods which are based on the homography and the neural network, respectively, for estimating the 3D-position of the hand from a camera image.

2.1 Hands Detection

The color based object recognition is used to detect the hand in a camera image. A red colored glove is used in this research to help distinguishing the hand from the other objects in environment. Fig. 1 shows the binary image with the number of pixels of detected hand. The size of an object in the image has the information about the distance on z-axis if it is spherical like a ball. Even though the hand is not spherical, the size of the hand is still useful information to find the distance on z-axis in a 2D-image.



Fig. 1. The result of detected hand

2.2 Homography-Based Method

Homography has been used in the field of computer vision. Any two images of the same planar surface in space are related by homography. If there is proper initial processing to find out the relation between the distorted 3D-position in the camera image and the 3D-position in the real world, the homography matrix can be calculated. The eight positions are sampled for initialization, which is described in the following.

1. The user imagines a hexahedral available space for the hand gesture interface.
2. The user moves the hand to each corner of the available space and inputs the corresponding position information one by one, as shown in Fig. 2.
3. Calculating the homography matrix between the image space and the real world by using singular value decomposition:

$$\begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = H \begin{bmatrix} u \\ v \\ w \\ 1 \end{bmatrix} \quad (1)$$

where H is the homography matrix, (u, v) is the position in the image space, w is the size of the hand in the image space, and (x, y, z) is the position in the real world.

Fig. 3 shows eight input positions and sizes of the hand in the image space according to each marker. The radius of each circle is the same as the square root of the number of pixels of detected hand.

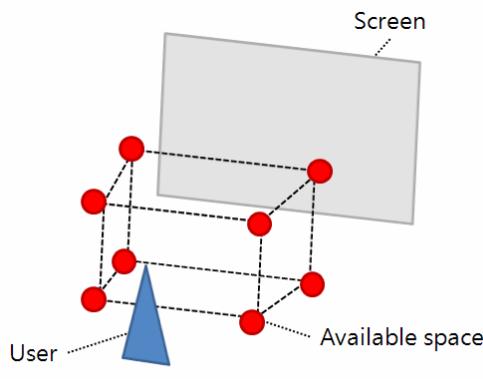


Fig. 2. Initialization for homography method

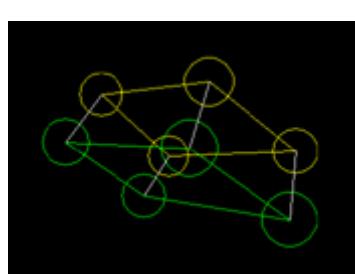


Fig. 3. The input data for homography method

2.3 Neural Network-Based Method

The size of the hand in the camera image does not guarantee the depth information because the hand is not spherical. The pose and the size of the hand in the camera image are almost the same when it returns to the same position while a user moves it freely. The neural network can be trained by the relation between the distorted 3D-position in the camera image and the 3D-position in real world. The initialization process is described in the following.

1. The user imagines a hexahedral available space for the hand gesture interface.
2. The initialization program shows a moving marker on the screen as shown in Fig. 4. This is the target data of the neural network. The user moves the hand continuously to proper positions according to the marker. This is the input data of the neural network (It takes about 15~25sec.).
3. Training the neural network (It takes about 10~15sec.).

Fig. 5 shows the input positions and sizes of the hand in the image space according to the moving marker.

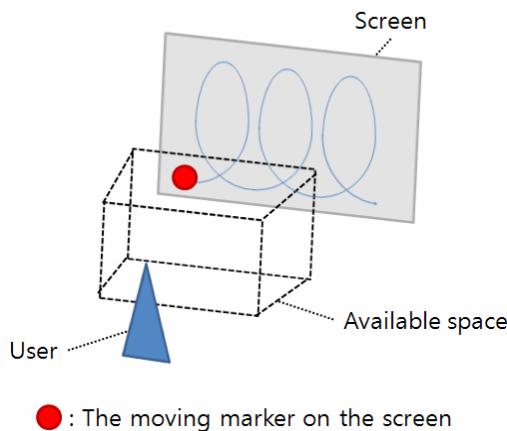


Fig. 4. Initialization for the neural network method

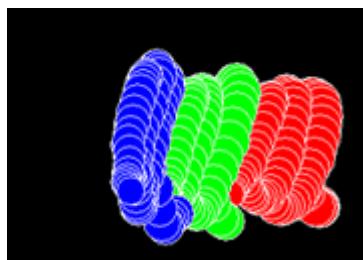


Fig. 5. The training data for the neural network method

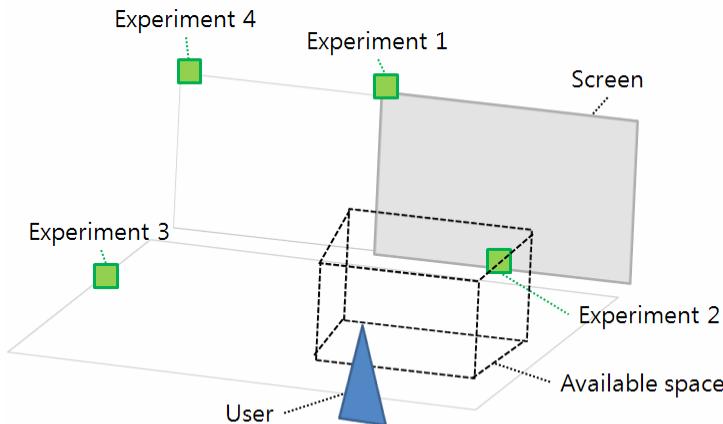
3 Experimental Results

3.1 Position Error in MS Windows Environment

Both methods were tested four times and the camera was relocated at each time as shown in Fig. 6. The test program drew 50 target positions on the screen one by one and each target position was generated randomly. It showed each target position during one second and users pointed the target position in their own interface space. It saved the estimated 3D-position of the hand when it changed the target position to the next one. The error is calculated by the distance between the target position and the estimated 3D-position of the hand in the x-y plane only, since it was tested in the MS windows environment. The position space was normalized [-1,1] on each axis. Table 1 shows the average errors and the standard deviations of the experiments.

Table 1. The experimental results

Experiment #	1	2	3	4
Homography	AVG	0.1106	0.0764	0.0797
	STD	0.4317	0.1157	0.1948
	%	21.59	5.78	9.74
Neural network	AVG	0.0917	0.0839	0.0862
	STD	0.2318	0.2947	0.1504
	%	11.59	14.73	7.52
				8.84



■ : The camera positions for the experiments

Fig. 6. The experimental environment

The method using the neural network generally showed better performance than the one using homography. The one using homography was sensitive to the location of the camera because the hand is not spherical. In summary, the method using the

neural network is better to estimate the 3D-position of the hand by using a single camera. Exceptionally, the method using homography is more efficient than the one using the neural network when the camera is located in front of the user or a spherical object is used instead of a hand.

3.2 Applying to Mouse Gesture Recognition Software

The proposed methods were tested with the Stroke-It; a mouse gesture recognition software [6]. If the position value of the hand on the z-axis was lower than the predefined threshold, it was interpreted as a pressed down mouse button (Fig. 7). The user held down the mouse button by a hand gesture and then drew the gesture. Once the gesture was performed, the software would execute the action associated with the gesture. The mouse gesture recognition software executed the web browser when the mouse drew ‘w’ (Fig. 8).

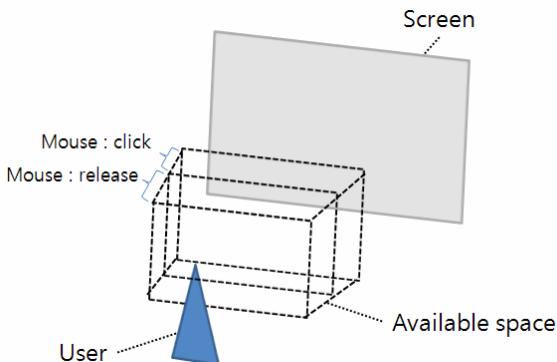


Fig. 7. The experiment to control the cursor of the mouse

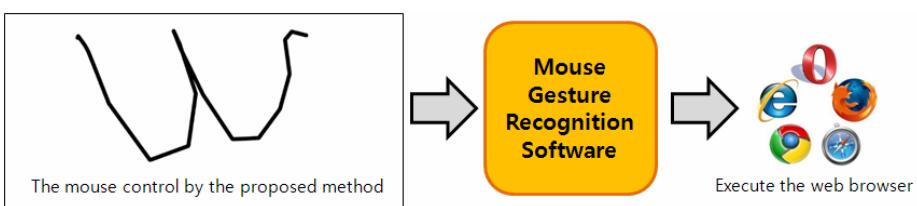


Fig. 8. The experiment with the mouse gesture recognition software

4 Conclusion

This paper dealt with the 3D-position estimation of the hand using a single camera. The two methods using homography and the neural network were proposed to estimate the 3D-position of the hand from a camera image. They were both implemented and tested with the camera at not only in front of the user but also

various locations. The experiment with the mouse gesture recognition software showed that the proposed method could be easily utilized in the office environment or in the presentation room. As the future work, the other information of the hand in a camera image should be applied to the neural network method to improve the performance.

Acknowledgments. This research was supported by the MKE (The Ministry of Knowledge Economy), Korea, under the National Robotics Research Center for Robot Intelligence Technology support program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2010-N02100128).

References

1. Hoshino, K., Tomida, M.: 3D Hand Pose Estimation Using a Single Camera for Unspecified Users. *Journal of Robotics and Mechatronics* 21(6), 749–757 (2009)
2. Ueda, E., Matsumoto, Y., Imai, M., Ogasawara, T.: Hand Pose Estimation for Vision-based Human Interface. *IEEE Transactions on Industrial Electronics* 50(4), 676–684 (2003)
3. Jeong, M.H., Kuno, Y., Shimada, N., Shirai, Y.: Recognition of Two-Hand Gesture Using Coupled Switching Linear Model. *IEICE Transactions on Information and Systems* E86-D(8), 1416–1425 (2003)
4. Athitos, V., Scarloff, S.: An Appearance-based Framework for 3D Hand Shape Classification and Camera Viewpoint Estimation. In: Proc. IEEE International Conference on Automatic Face and Gesture Recognition, Washington, DC, pp. 40–45 (2002)
5. Wu, Y., Lin, J., Huang, T.S.: Analyzing and Capturing Articulated Hand Motion in Image Sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27(12), 1910–1922 (2005)
6. StrokeIt, <http://www tcbmi com/strokeit/>

Hand Gesture for Taking Self Portrait

Shaowei Chu and Jiro Tanaka

Department of Computer Science, University of Tsukuba,
Tennodai, Tsukuba, 305-8577 Ibaraki, Japan
chushaowei@iplab.cs.tsukuba.ac.jp,
jiro@cs.tsukuba.ac.jp

Abstract. We present a new interaction technique enabling user to manipulate digital camera when taking self-portrait pictures. User can control camera's functions such as pan, tilt, and shutter using hand gesture. The preview of camera and GUIs are shown on a large display. We developed two interaction techniques. First one is a hover button that triggers camera's shutter. Second one is cross motion interface that controls pan and tilt. In this paper, we explain algorithms in detailed manner and show the preliminary experiment for evaluating speed and accuracy of our implementation. Finally, we discuss promising applications using proposed technique.

Keywords: Hand gesture, self-portrait, human computer interaction, skin color, hand detection, fingertip detection, optical-flow.

1 Introduction

Digital camera is a way to take self-portraits allowing us to express ourselves through photography. Nowadays digital cameras have many functions to take self-portrait pictures, such as frontal screen [6], self-timer, face and smile detection [6], motion detection [7]. However, the main drawback of these methods is that it is hard to control the camera from a distance. Moreover the preview is too small to see from a distance. Therefore, when the user sets up the shooting parameters and timer, runs into front to the camera, while it is impossible to change the camera settings or know whether or not they are in the good view to shoot. For such reasons a common scenario is one where user runs back and forward to make sure a good shot is obtained. Besides, when using SLR (single lens reflection) camera, in case of a little bit wrong position, nearer or farther from the camera, the result image will be blurred because the lens focal.

In this paper, we explore the hand tracking and hand motion gesture recognition approach to interact with the camera for taking self-portrait pictures. We apply computer vision algorithm on camera live view video to recognize user's hand gestures. The pan and tilt platform guarantees the camera will focus on the user automatically. The large display shows the augmented live view video, which gives a clear preview to user. By using hand gestures, the user can control the camera functions from a distance. Figure 1 shows the system overview.



Fig. 1. The system overview (left), setup scene (middle), and augmented live view (right)

2 Related Work

Many innovative techniques have been proposed in the literature to deal with the difficulties in computer vision to control the devices from a distance. Vision based hand gesture recognition is believed to be an effective technique [3]. Therefore, a number of systems have been proposed.

Chen [4] presents an optical flow with MoSIFT appearance features, to recognize gestures for controlling TV operations. The MoSIFT is computationally expensive, the authors' implementation is based on parallel processing with multi-core processors to improve the recognize latency. However, the result shows that it still requires quiet long processing time. It takes about 2.5 seconds between an activity and a result action.

Lenman [5] shows a study of using gestures interact with pie and marking menus as a remote controller to control electronic appliances in a home environment, such as TV sets and DVD players. The system recognizes the hand poses based on a combination of multi-scale color feature detection, view-based hierarchical hand models and particle filtering. The hand poses can be detected then tracking the movement. The problem of this approach is that it is slow and not accurate on complex background. Moreover the user needs to hold the hand pose still and frontally facing the camera for several seconds to activate an operation.

Sixth Sense [6] uses colored marker attached to fingertips and Microsoft Kinect uses depth image sensor to segment people's body, both of their approaches make the hand detection more easier. However we want to pursuit a pure vision based method that is marker-less and can easily embedded into everyday used common architecture digital cameras.

In short, most of the presented interaction techniques of hand gestures interaction have limitations for the purpose of taking self-portraits by using a digital camera. Our proposed interaction technique is application oriented designed especially for self-portrait; a higher speed hand detecting algorithm and a cross motion recognition interface are developed. By using this light-weighted algorithm, it is easier to transfer the algorithm into the camera device.

Our proposal mainly has three contributions. First, we propose a novel technique that enables user to manipulate digital camera conveniently using hand gesture

especially when controlling it from a distant. Second, we developed a real-time computer vision algorithm that tracks the hand and fingertip with accuracy and high speed. Third, a cross motion interface used to recognize hand motion direction has been proposed.

3 System Overview

In this section we will discuss the hardware we used, and low-level image processing algorithm to detect the hand and hand motion gestures.

3.1 Hardware

The camera we used is Logitech Orbit AF Camera, capturing 640 x 480 resolution video with 30 FPS, 1024 x 768 for 15 FPS and 1600 x 1200 for 5 FPS, with 189-degree field of pan and 102-degree field of tilt.

The large display we used has a size of 30 inches which will give a clear augmented live view from a distance. We use down sized 320 x 240 resolution live view video applies computer vision algorithm, 640 x 480 resolution for live view show and 1600 x 1200 resolution for taking shots.

3.2 Hand Contour Segmentation and Fingertip Detection

The hand tracking applies low-level image processing operations on each frame of video in order to detect the locations of the fingers. While a model-based [2] approach that uses temporal information could provide more robustness to situations such as complex backgrounds, has been implemented. Our algorithm of hand detection is described in the following:

Given a captured image, every pixel is categorized to be either a skin-color pixel or a non-skin-color pixel. We use skin-color detection algorithm, described in [1], to segment the skin-color pixels. After segmenting, an amount of noise pixels in the image is inevitable, we apply median filtering [8] with 3 x 3 kernel to remove extraneous noise. Then we apply single connected component contour finding algorithm (implemented in OpenCV [9]) to locate hand contours. We abandon the small area contours, which are less than 600 pixels, and leave only larger ones as hand candidates.

After we obtain the contours, the next job is to detect the fingertips. The fingertips are detected based on the contour information by using a curvature-based algorithm similar to the one described in [2]. The algorithm can detect the fingertips and finger orientation base on 2D hand silhouette. Figure 2 shows the processing of hand fingertips detection from the original image to hand fingertips detected result image. During this process, because the color of the face is also skin-color, we have detected the face as well; this will be useful for face detection and recognition in our future work.

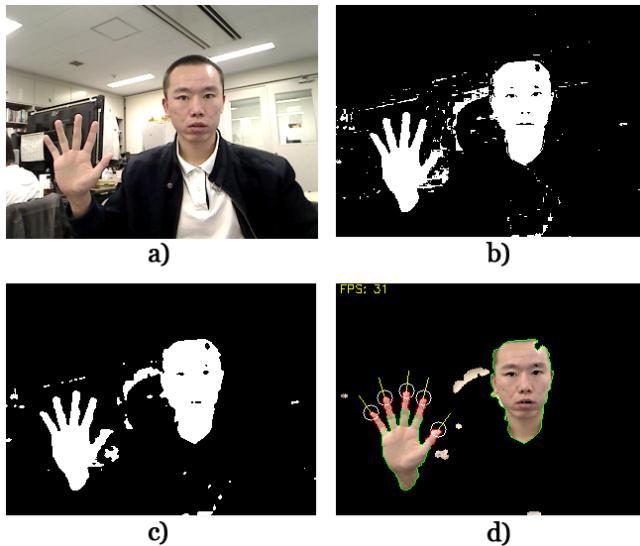


Fig. 2. Hand fingertips detection procedure. a) original image. b) skin-color segment. c) median filtered. d) detected fingertips with colored marker.

3.3 Motion Gesture Recognition

In this section we will describe a motion recognition interface designed for recognizing 4 motion direction gestures, wave UP, DOWN, LEFT and RIGHT, by using the optical-flow measurements. The motivation of developing such interface is that when the user moves his/her hand with a certain speed, the hand image will blur and it is very difficult to detect the fingertips. The optical-flow [9] denotes the movement of an image patch between two frames of a video sequence that can measure the motion gesture even with high speed movement which is suitable for our scenario. There are various techniques for estimating the optical-flow [11] [12], and it is proved efficient [10] for gesture recognition.

Calculating the optical-flow in real-time for the whole image at 320 x 240 resolution might require a lot of computing power. We restrict the optical-flow measurement with limited feature points (29 points) and within a small region. The feature points movement will be extracted in each frame and calculating the mean value of both orientation and speed. The noise of small and large movement of feature points will be cut off, leaving only reliable movement among them. In order to recognize the gestures we analyse the pattern of feature points movement in video frame sequence, from no movement to movement, and to no movement, then distinguish a specific motion gesture.

Optical-flow estimates are often very noisy; we add 29 feature points, and calculate the main motion measurements using mean orientation value, the result shows it is increase to indicate the expected gesture by adding more points. The layout of feature points affects the optical-flow measurement also; we apply a circle-like layout which proves effective for recognizing four motion direction. Figure 3 shows a frame sequence in video time line, that to recognizing a RIGHT hand motion.

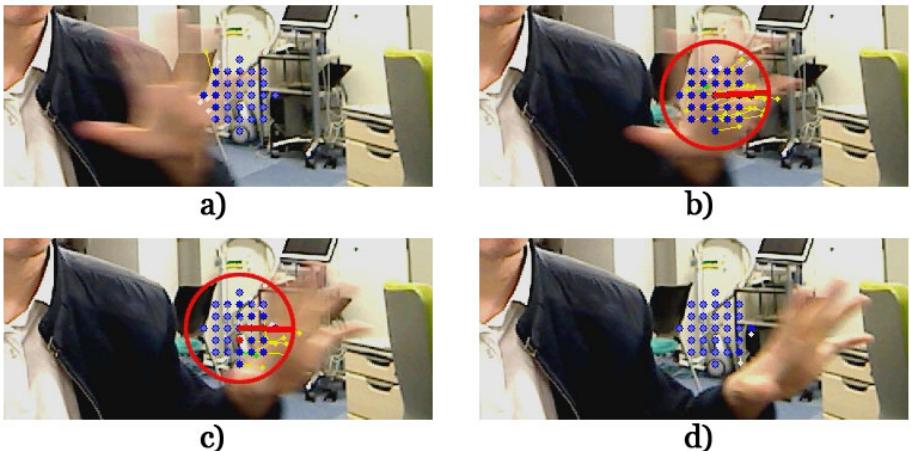


Fig. 3. The procedure of motion gesture frame time line. a) frame t, no motion detected. b) frame t + 1, motion orientation detected, the clock and arrow indicate the motion orientation is right. c) frame t + 2, motion orientation is right. d) frame t + 3, no motion detected.

4 Interaction Techniques

We use the large display to give augmented live view, including the highlighted color marker of hand, fingertips and graphical interface, as in Figure 4. Two interfaces have been implemented.

4.1 Hover Button Interface

The hover button is a fingertip air touch interface to activate the camera shutter. On the large display the detected hand will be marked with green contour pixels, the fingertips will be highlighted with red pixels. User can put out any fingertip into the region of shutter button to activate the button. Because of the noise of detection and cluster background, (s)he needs to put the fingertip into the region of button steady for a specific period, one second, to activate the shutter button. After the shutter is activated, a big colored number on large display will be shown counting down from 3 to 1, which indicates the user should prepare for a pose, and taking a shot (see Figure 1 b)).

4.2 Cross Motion Interface

The second interface is a motion based interface as described in section 3.3, which is used to adjust the camera's pan and tilt. We designed a GUI for cross motion interface (see Figure 4, Cross motion interface). The four direction arrows indicate it can recognize four motion directions, pan left, right and tilt up, down. When user uses a hand to make a cross motion within a specific short period cross the interface, then the hand motion direction, UP, DOWN, LEFT and RIGHT, can be recognized by using optical-flow measurements described in section 3.3. The camera pan & tilt

operations will function according to the hand motion direction. To make the interface more robust, which prevents the camera from moving in a disorderly manner, the gestures of slow motion and move continually within the interface region will be excluded. For 20 milliseconds after performing a motion, the user needs to stand still (no motion on the interface). After that, a pattern of motion will be recognized and the pan & tilt operations will be functional immediately. Figure 3 shows one of RIGHT motion gesture. Figure 4 shows the augmented live view image including two GUIs.

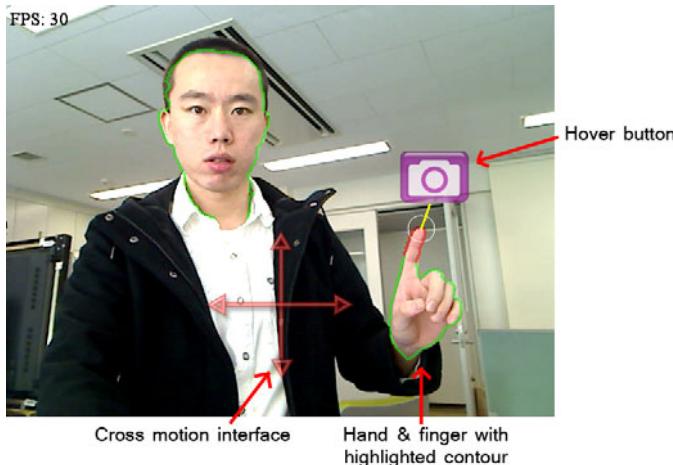


Fig. 4. Augmented live view, mixes with colored hand marker and GUIs

5 Performance Evaluation

5.1 Performance of Hand Fingertip Detection and Optical-Flow

We evaluated two algorithms with regard to speed and accuracy. The experiments were performed on a computer with a 2.5GHz CPU, using USB 2.0 Pan & Tilt camera.

Table 1 shows the process time of hand fingertip detection. The result shows that the hand fingertip detection algorithm can run in real-time with about 180 FPS. Optical-flow measurement of cross motion interface is less than 3 milliseconds.

Table 1. The process time of hand fingertip detection with video resolution at 320 x 240

Processing time	Process time (millisecond)
Skin-color segmentation	0.6
Mean Filter	3.7
Finding Contour	0.5
Fingertip Detection	0.7
Total	5.5

5.2 Accuracy of Hand Tracking and Cross Motion Recognition

The goal of this evaluation is to test the distance where the hand tracking and cross motion interface work well.

For testing the hand tracking, we asked the users to put the hand out in front of the camera to see whether or not the hand can be detected. By various people test the algorithm works well from 0.5 meter to 2 meters away from camera. For testing the accuracy of cross motion interface, we asked users to perform gestures and to see whether they are correctly recognized. Ten volunteers participated in the experiment. The results are shown in tables 2, 3 and 4. Left column of the table means user performed gestures, corresponding row is result of recognized gestures, on average percentage.

Table 2. Cross motion interface accuracy (distance 0.5 meter)

Gesture	LEFT	RIGHT	UP	DOWN	NONE
LEFT	90%				10%
RIGHT		84%		3%	13%
UP	7%		90%		3%
DOWN		4%		90%	6%

Table 3. Cross motion interface accuracy (distance 1.2 meters)

Gesture	LEFT	RIGHT	UP	DOWN	NONE
LEFT	80%			2%	18%
RIGHT		82%			18%
UP			86%	3%	11%
DOWN			2%	83%	15%

Table 4. Cross motion interface accuracy (distance 2 meters)

Gesture	LEFT	RIGHT	UP	DOWN	NONE
LEFT	52%				48%
RIGHT		64%			36%
UP			85%	1%	14%
DOWN			3%	88%	9%

6 Discussion

6.1 Limitations

We have shown the accuracy and effectiveness of our proposed gesture interfaces. However a few limitations do exist.

First, lighting sensitivity is a main weakness for skin-color based hand detection. Outdoors or in highly reflective sunshine conditions, the human skin-color will

change to white color, what will cause difficulty to segment. Although many researchers have used uniform background segment [2], color histogram [15], flocks of features [13] to improve the hand tracking, these methods are either slow or less accurate. Therefore we choose another method to adjust camera optics parameters, Exposure, Gain, Brightness, Contrast, Color Intensity, etc., which maintains the video image with uniform balanced lighting condition. Moreover, fortunately, modern camera have self-adaptive functions to adjust them automatically or manually.

Second, the distance from the user to the camera is another issue; the user needs to stand within 2 meters from the camera, otherwise it will be difficult to detect the hand. The cross motion interface also has this problem.

Third, the optical-flow measurements need a cluster background; if the image region on the optical-flow cross motion interface is a completely uniform color, the optical-flow measurements will inaccurate. Such situations occur when the user worn a uniformly colored clothes.

6.2 Possible Applications

Our proposed two gesture based interface can be applied to various applications.

The hand and finger detection algorithm can be widely used in vision based human computer interaction. For example, for HMD and Augmented Reality, the hand gesture recognition is very practical to use for interacting with virtual objects and interface. The finger orientation can do distance pointing as described in [16]. Two hand gestures interaction can also be developed (See Figure 5).



Fig. 5. Finger orientation interface and two hand manipulation of picture

The motion based gesture recognition is also a novel approach to recognize the human hand gestures. All of the gestures represent a sequence of motion in video. By carefully analyzing the motion frame sequence in video time line, specified gesture can be recognized. We developed a motion based interface and introduced it into our application for controlling the pan and tilt, and we obtained good results. However, this is just a primary attempt, numerous interface can be implemented, such as the slider bar, clock move interface, button. Moreover, if we set the interface on the region of the face, then the head shake and nod gesture can be recognized as well.

7 Conclusions and Future Work

In this paper we presented the hand gestures for taking self-portraits. The camera's pan, tilt and shutter can be controlled by hand gestures. We use skin-color and k-curvature algorithm to detect the fingertip, optical-flow measurement to recognize the hand motion gesture, both of them running with accuracy and high speed. By using our system, the user can control the self-portrait process from a distance. Our result showed that hand gesture is promising interface as a means of manipulating digital camera remotely.

In the future, we have plan to support more functionalities with a SLR camera and a more accurate Pan and Tilt platform for our experiment. We also want to apply face tracking and recognition techniques to self-portrait system which will be benefit for user tracking and multi-users scenario. Finally, new interaction techniques that can work outdoors, with small or screen-less live view cameras, have also come into consideration.

References

1. Kovač, J., Peer, P., Solina, F.: Human Skin Colour Clustering for Face Detection. In: IEEE EUROCON 2003: International Conference on Computer as a Tool, vol. 2, pp. 144–148 (2003)
2. Malik, S., Laszlo, J.: Visual Touchpad: A Two-handed Gestural Input Device. In: The ACM International Conference on Multimodal Interfaces, pp. 289–296 (2004)
3. Mistry, P., Maes, P.: SixthSense: a wearable gestural interface. In: ACM SIGGRAPH ASIA 2009 Sketches, pp. 11:1–11:1 (2009)
4. Chen, M.Y., Mummert, L., Pillai, P., Hauptmann, A., Sukthankar, R.: Controlling Your TV with Gestures. In: MIR 2010: 11th ACM SIGMM International Conference on Multimedia Information Retrieval, pp. 405–408 (2010)
5. Lenman, S., Bretzner, L., Thuresson, B.: Using Marking Menus to Develop Command Sets for Computer Vision Based Hand Gesture Interfaces. In: Proceedings of the Second Nordic Conference on Human-computer Interaction, pp. 239–242 (2002)
6. Sony Party Shot (2009), <http://www.sony.jp/cyber-shot/party-shot/>
7. Samsung Self-Portrait Camera: TL225 (2009),
<http://www.samsung.com/us/photography/digital-cameras/>
8. Pereault, S., Hebert, P.: Median Filtering in Constant Time. IEEE Transactions on Image Processing 16, 2389–2394 (2007)
9. OpenCV: Open Source Computer Vision,
<http://opencv.willowgarage.com/wiki/>
10. Zivkovic, Z.: Optical-flow-driven Gadgets for Gaming User Interface. In: International Conference on Entertainment Computing, pp. 90–100 (2004)
11. Lucas, B., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: 7th International Joint Conference on Artificial Intelligence, vol. 2, pp. 674–679 (1981)
12. Beauchemin, S.S., Barron, J.L.: The Computation of Optical Flow. ACM Computing Surveys, 433–467 (1995)
13. Kölisch, M., Turk, M.: Fast 2D hand tracking with flocks of features and multi-cue integration. In: Conference on Computer Vision and Pattern Recognition Workshop, vol. 10, p. 158 (2004)

14. Letessier, J., Bérard, F.: Visual tracking of bare fingers for interactive surfaces. In: Proceedings of the 17th Annual ACM Symposium on User Interface Software and Technology (UIST 2004), pp. 119–122 (2004)
15. Lee, T., Höllerer, T.: Handy AR: Markerless Inspection of Augmented Reality Objects Using Fingertip Tracking. In: International Symposium on Wearable Computers (IEEE ISWC), pp. 11–13 (2007)
16. Wang, F., Cao, X., Ren, X., Irani, P.: Detecting and Leveraging Finger Orientation for Interaction with Direct-Touch Surfaces. In: Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology (UIST 2009), pp. 23–32 (2009)

Hidden-Markov-Model-Based Hand Gesture Recognition Techniques Used for a Human-Robot Interaction System

Chin-Shyurng Fahn and Keng-Yu Chu

Department of Computer Science and Information Engineering
National Taiwan University of Science and Technology

Taipei, Taiwan 10607, Republic of China
csfahn@mail.ntust.edu.tw

Abstract. In this paper, we present part of a human-robot interaction system that recognizes meaningful gestures composed of continuous hand motions in real time based on hidden Markov models. This system acting as an interface is used for humans making various kinds of hand gestures to issue specific commands for conducting robots. To accomplish this, we define four basic types of directive gestures made by a single hand, which are moving upward, downward, leftward, and rightward individually. They serve as fundamental conducting gestures. Thus, if another hand is incorporated to making gestures, there are at most twenty-four kinds of compound gestures by the combination of the directive gestures using both hands. At present, we prescribe eight kinds of compound gestures employed in our developed human-robot interaction system, each of which is assigned a motion or functional control command, including moving forward, moving backward, turning left, turning right, stop, robot following, robot waiting, and ready, so that users can easily operate an autonomous robot. Experimental results reveal that our system can achieve an average gesture recognition rate of 96% at least. It is very satisfactory and encouraged.

Keywords: hand gesture recognition, hidden Markov model, human-robot interaction, directive gesture, compound gesture.

1 Introduction

In order to realize a human computer interface that possesses more human nature and is easy to operate, the researches focusing on hand gesture recognition have been rapidly grown up to make humans easier to communicate and interact. Nowadays, the control panel is gradually switched from a keyboard to a simple hand touch, which dramatically changes the communication style between humans and computers. Regarding the hand gesture recognition, it also has many applications such as combining face and hand tracking for sign language recognition [1], using fingers as pointers for selecting options from a menu [2], and interacting with a computer by an easy way for children [3]. Over the last few years, many methods for hand gesture recognition were proposed. Compared with face recognition, hand gesture recognition has different problems to be overcome. First of all, the main features of hand gestures

are obtained from the fingers and palms of our hands. Each finger has three joints and the palm of a hand connects with the wrist. Therefore, the hand gestures can produce many great changes on the postures, especially for the different form of changes resulting from varied hand gestures. In addition, the hand gestures can have three-dimensional spatial revolution. Consequently, the hand gesture recognition may not be easier than the face recognition.

Vision-based gesture analysis has been studied for providing an alternative interaction between humans and computers or robots in recent year. In particular, the hand gesture recognition has become a major research topic for a natural human-robot interaction. Users generally exploit arms and hand gestures to give expression of their feelings and notification of their thoughts. And users also employ more simple hand gestures such as pointing gestures or command gestures rather than complex gestures. Kim et al. [4] proposed vision-based gesture analysis for human-robot interaction that includes the detection of human faces and moving hands as well as hand gesture recognition. Their method for detection is resorted to skin colors and motion information, and for recognition is to adopt neural networks. The main disadvantage is the user must wear a long-sleeves shirt that only the skin color of a palm is exposed. Our research is to conduct a robot directly by hand gestures without any auxiliary wearable or special equipment. To implement this, we intend to develop a human-robot interaction system to recognize some gestures defined by users via a PTZ camera capturing color image sequences in real time.

In this paper, part of the human-robot interaction system installed on an autonomous robot is presented. This system consists of four main processing stages: face detection and tracking, hand detection and tracking, feature extraction, and gesture recognition. Prior to the process of gesture recognition, we devise an automatic hand localization method based on skin colors and circle detection for finding palms, which is quite robust and reliable to complete hand detection and tracking in unconstrained environments; thus, the extraction of hand regions is very fast and accurately. Such a hand localization method may not be confined to uncover the lower arm, so users are not required to wear long-sleeves shirts. With the aid of hand localization to achieve hand detection and tracking, the orientation of a hand moving path between two consecutive points that stand for the positions of a palm appearing in a pair of image frames is extracted as an essential feature for gesture recognition. Subsequently, by virtue of the classification scheme adopting hidden Markov models (HMMs) that are widely applied to speech or handwriting recognition, we can effectively acquire the result of hand gesture recognition and issue the corresponding command to conduct the robot as expected.

2 Feature Extraction of Hand Gestures

2.1 Hand Detection and Tracking

To realize our human-robot interaction system that can automatically recognizes hand gestures consisting of continuous hand motions, hand localization (a single hand or two hands) is a crucial process because the information of hand regions rather than a face region are what we want. The hand localization process includes hand detection,

tracking, and feature extraction. Due to the limitation of article length, the detailed description of the hand detection and tracking can refer to our past research [5].

2.2 Feature Extraction

What follows introduces a method to analyze the features of a hand gesture and the ways to extract them before recognition. There is no doubt that appropriate feature selection to classify a hand gesture plays a significant role of improving system performance. Herein, three basic features: location, orientation, and velocity are adopted. From the literature [6, 7], they showed that the orientation feature is the best one to attain high accurate recognition results. Hence, we will treat the orientation as a main feature used in our system to recognize a meaningful gesture composed of continuous hand motions constituting a hand motion trajectory.

The hand motion trajectory is a spatio-temporal pattern that can be represented with a sequence of the centroid points $(x_{\text{hand}}, y_{\text{hand}})$ of detected hand regions. Consequently, the orientation between two consecutive points in the hand motion trajectory is determined by Equation (1).

$$\theta_t = \arctan \left(\frac{y_{t+1} - y_t}{x_{t+1} - x_t} \right) \quad t = 1, 2, \dots, T-1 \quad (1)$$

where T represents the length of a hand motion trajectory. The orientation is quantized into 12 levels, each of which is separated by 30° in order to generate twelve direction codewords called 1 to 12 as Figure 1 shows. Therefore, a discrete vector is in form of a series of direction codewords and then used as the input to an HMM.

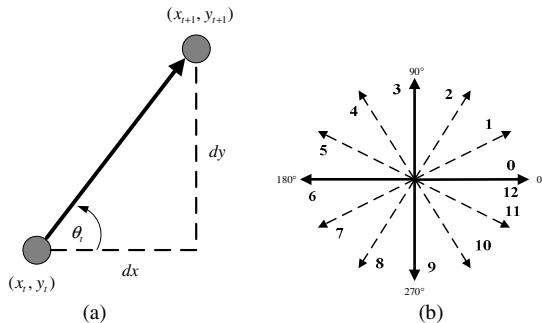


Fig. 1. The orientation and its quantization levels: (a) the orientation between two consecutive points; (b) the twelve quantization levels where direction codeword 0 is equivalent to direction codeword 12

3 Gesture Recognition

Gesture recognition is a challenging task for distinguishing a meaningful gesture from the others. This section depicts our human-robot interaction system that humans make various kinds of hand gestures to issue specific commands for directing robots. To

achieve this, we define four basic types of directive gestures made by one hand, which are moving upward, downward, leftward, and rightward individually. They serve as fundamental conducting gestures. Thus, if another hand is incorporated to making gestures, there are at most twenty-four kinds of compound gestures by the combination of directive gestures using both hands. Finally, at the gesture recognition stage in our human-robot interaction system, we will apply a probabilistic approach, the HMM, to classify a hand gesture as shown in Figure 2.

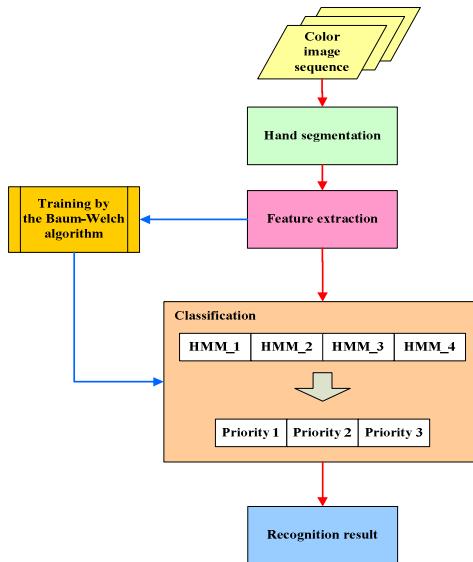


Fig. 2. The flow chart of the gesture recognition procedure

3.1 Gesture Definition

Issuing commands to direct robots through only vision without sounds or other senses is similar to conducting a marching band by means of visible gestures. In order to accomplish real-time operations, our system requires simpler body language which is easy to recognize and discriminate from each other. Very many kinds of daily used gestures are characterized by hand motions. In this system, four basic types of directive gestures: moving upward, downward, leftward, and rightward made by one hand are defined to serve as fundamental conducting gestures. By the combination of directive gestures using both hands simultaneously, we will have at most twenty-four kinds of compound gestures. For convenience' sake, a 2-D table is employed to express all the meaningful gestures, and each of them is named a gesture ID respectively. In this manner, it easily symbolizes every gesture and is convenient to add new hand gestures.

In essential, the gesture recognition techniques applied to single right or left hand motions are all the same. For practical use, we must continue to distinguish which one of the two hands is making gestures. At present, eight kinds of meaningful gestures are prescribed for directing robots. The gesture usage contains motion control and

functional control. Each of them is assigned a command to conduct robots in our system, which includes moving forward, moving backward, turning left, turning right, stop, robot following, robot waiting, and ready. Table 1 illustrates the above gestures and lists their associated right and left hands features in terms of orientation sequences, respectively. Notice that in order to interact with an autonomous robot smoothly, we specially define one of the eight kinds of gestures, which both hands are put in front of the breast, as a static gesture to make stopping of any robot motion while the robot is operating. According to the features extracted from hand motion trajectories, we feed them to an HMM-based classifier for training and recognizing a given hand gesture seen from the human-robot interaction system.

Table 1. Corresponding Commands of Eight Kinds of Gestures

Command	Gesture	Left hand feature	Right hand feature
Moving forward		Orientation sequence = NULL	Orientation sequence = [10, 10, 10, 10, 10, 10]
Moving backward		Orientation sequence = [1, 1, 1, 1, 1, 1]	Orientation sequence = [7, 7, 7, 7, 7, 7]
Turning left		Orientation sequence = [1, 1, 1, 1, 1, 1]	Orientation sequence = NULL
Turning right		Orientation sequence = NULL	Orientation sequence = [7, 7, 7, 7, 7, 7]
Stop		Orientation sequence = NULL	Orientation sequence = NULL
Following		Orientation sequence = [10, 10, 10, 10, 10, 10]	Orientation sequence = [10, 10, 10, 10, 10, 10]
Waiting		Orientation sequence = [10, 10, 10, 10, 10, 10]	Orientation sequence = [7, 7, 7, 7, 7, 7]
Ready		Orientation sequence = [1, 1, 1, 1, 1, 1]	Orientation sequence = [10, 10, 10, 10, 10, 10]

3.2 The Hidden Markov Model

HMMs are chosen to classify the gestures, and their parameters are learned from the training data. Based on the most likely performance criterion, the gestures can be recognized by evaluating the trained HMMs. However, the HMM is different from the Markov model. The latter is a model with each state corresponding to an observable event and its state transition probability depends on both the current state and predecessor state. And extending from the Markov model, the HMM considers more conditions of the observable event, so it has been excessively used in various fields of applications such as speech recognition [8] and handwriting recognition [9]. Because the HMM is more feasible than the Markov model, we adopt the former to learn and recognize the continuous hand gestures to direct robots.

3.3 Fundamentals of HMMs

An HMM consists of a number of states, each of which is assigned a probability of transition from one state to another state. Additionally, each state at any time depends only on the state at the preceding time. In an HMM, one state is described by two sets of probabilities: one is composed of transition probabilities, and the other is of either discrete output probability distributions or continuous output probability density functions. However, the states of an HMM are not directly observable, but they can be observed through a sequence of observation symbols. Herein, we bring in the formal definition of an HMM [10] which is characterized by three matrices: the state transition probability matrix A , symbol output probability matrix B , and initial state probability matrix π . They are all determined during the training process and simply expressed in a set of $\lambda = \{\pi, A, B\}$. Figure 3 illustrates a state transition diagram of an HMM with three hidden states and three observations.

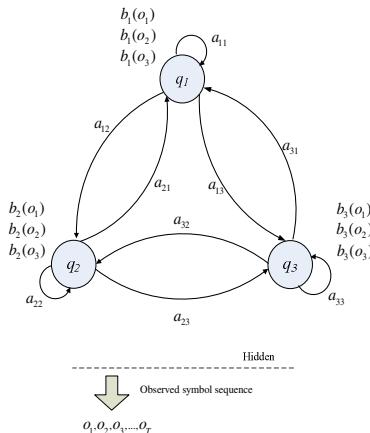


Fig. 3. Illustration of an HMM with three hidden states and three observations

Three states are drawn as circles in this example. Besides, each directed line is a transition from one state to another, where the transition probability from state q_i to state q_j is indicated by a_{ij} . Note that there are also transition paths from states to themselves. These paths provide the HMM with time-scale invariations because they allow the HMM to stay in the same state for any duration. Each state of the HMM stochastically outputs an observation symbol. In state q_i , for instance, symbol o_k is the output with a probability of $b_i(o_k)$ at time step k . During a period, the state yields a symbol sequence $O = \{o_1, o_2, \dots, o_T\}$ from time step 1 to T . Thus, we can observe the symbol sequences output by the HMM, but we are unable to observe its states. In addition to this, the initial state of the HMM is determined by the initial state probability $\pi = \{\pi_i\}, 1 \leq i \leq N$.

In general, three tasks: evaluation, decoding, and training should be accomplished under the framework of an HMM. In practice, they can be commonly solved by the forward-backward algorithm, Viterbi algorithm, and Baum-Welch algorithm, respectively [10].

4 Gesture Recognition

At the gesture recognition stage, the Baum-Welch algorithm is used for training the initialized parameters of an HMM to provide the trained parameters. After the training procedure, both the trained parameters and the discrete vector derived from a hand motion trajectory are fed to the Viterbi algorithm to obtain the best hand moving path. Using this best path and a gesture database, the robot can recognize a given hand gesture made by users. The number of states in our realized HMM is based on the complexity of each hand gesture and is determined by mapping each straight-line segment into one state of the HMM as graphically shown in Figure 4. There are four basic types of directive gestures for interacting with robots, including moving upward, downward, leftward, and rightward individually. In addition, we convert a hand motion trajectory into an orientation sequence of codewords from 1 to 12 acquired at the feature extraction stage depicted in Section 2.2. In the HMM, the number of hidden states is set to 1 and the number of observation symbols is fixed to 12.

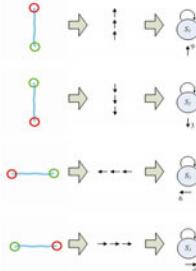


Fig. 4. Each straight-line segment corresponding to a type of directive gestures for establishing the topology of an HMM

Furthermore, a good parameter initialization of $\lambda = \{\pi, A, B\}$ for an HMM will produce better results. First, we determine the initial matrix A is $A = \{a_{11} = 1\}$ since for all four types of directive gestures in our system only contain one straight-line segment, each of which needs one state. Second, the parameter matrix B is evaluated by Equation (2). Because the states of an HMM are discrete, all elements of matrix B can be initialized with the same value for each of different states.

$$B = \{b_{im}\} \text{ with } b_{im} = \frac{1}{M} \quad (2)$$

where i is the serial number of states, m is the serial number of observation symbols, and M is the total number of observation symbols. Finally, the initial state probability π_i is defaulted by 1.

Once the parameters of an HMM are initialized, we utilize the Baum-Welch algorithm to perform the training process where the inputs of this algorithm are the initialized parameters and the discrete vector that is obtained from the feature extraction stage. Consequently, we will acquire the new parameters from such a training phase. In our system, the gesture database contains 20 sample videos, 15 of which are used for training and 5 for testing, including the four types of directive gestures. And we take 10 discrete vectors for each type of directive gestures. While the training process for each video is finished, we will take the discrete vector and the new parameters of the HMM as the inputs for the Viterbi algorithm. Therefore, we can get the best hand moving path that is corresponding to the maximal likelihood of the four types of directive gestures. In the sequel, we compare and choose the higher priority form the gesture database, then output the result of recognition.

5 Experimental Results

5.1 Tests on the Gesture Recognition

Figure 5 shows some training samples of hand gestures which our system can recognize. For a convenience, Roman numerals (I to IX) are assigned to symbolize each kind of gestures. We can understand the reason why the accuracy rates of recognizing Gestures I, III, IV, and V are higher than those of recognizing other kinds of gestures. Especially for Gesture V, its accuracy rate of recognition is the highest, because the feature values of putting both hands in front of the breast are easier to discriminate from each other than those of stretching two hands simultaneously. The accuracy rate of recognizing Gesture II is not desirable, on account of the influence of luminance and different motion speeds with both hands. Since we make all training samples not only out of doors but also within doors, sometimes the hands are too close to a fluorescent light when we raise them up.

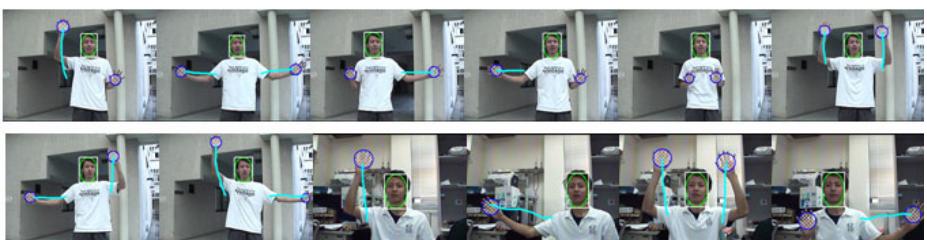


Fig. 5. Some training samples of gestures which our system can recognize

After calculating the necessary terms, we apply the recall and precision rates to measure the performance of our HMM-based hand gesture classifier. Table 2 lists the precision and recall rates of the experimental results. We can see that the recall rate measures how the proportion of a certain kind of subjects is classified into the correct kind of gestures, whereas the precision rate responds to the result of the

Table 2. The Recall and Precision Rates of the HMM-Based Hand Gesture Classifier

Kind of gestures \ System performance	Hidden Markov Models	
	Recall rate	Precision rate
Moving forward (I)	99.8%	97.1%
Moving backward (II)	97.0%	99.8%
Turning left (III)	99.4%	97.6%
Turning right (IV)	99.6%	97.8%
Stop (V)	100%	100%
Following (VI)	98.2%	100%
Waiting (VII)	98.8%	100%
Ready (VIII)	97.8%	100%
Others (IX)	98.8%	97.1%
Average	98.8%	98.8%

misclassification of some kinds of gestures. In other words, the higher precision rate means that the other kinds of gestures are rarely misclassified to the kind of the target gestures.

Table 3 shows the average accuracy rate of recognizing each kind of gestures using the HMM-based classifier. We can observe that the average accuracy rate of recognizing Gesture V is the highest by use of this classifier. The average accuracy rates of recognizing Gestures VI, VII, and VIII are not desirable, whose common characteristic is to raise one or two hands up. It can be inferred that the feature values of hands are not stable when they are raised up. We must find other feature values to enhance the stability to solve this problem. Most of gestures grouped into the kind of “Others” are to make both hands hang down naturally or cross around the torso and chest. Since the average accuracy rate of recognizing Gesture I is not bad, we consider adding new kinds of gestures in a similar way.

Table 3. The Average Accuracy Rate of Recognizing Each Kind of Gestures

Gesture Measurement \	I	II	III	IV	V	VI	VII	VIII	IX
Average accuracy rate	98.96%	98.40%	98.53%	98.76%	99.07%	96.07%	97.83%	96.93%	97.83%

5.2 Tests on the Human-Robot Interaction System

In this experiment, we equip our HMM-based hand gesture classifier on an autonomous robot actually. Through the eight kinds of gestures defined by us using both hands, we can interact with the robot easily. Such a kind of gestures is corresponding to a motion or functional control command; for example, we can raise the right hand up vertically, extend both hands horizontally, spread the left hand horizontally, and spread the right hand horizontally to conduct the robot moving forward, backward, leftward, and rightward, respectively. And, we can easily stop the current action of the robot by putting both hands in front of the breast. Besides, we devise another interaction mode that the robot follows a user. When the user raises his/her both hands simultaneously, the robot will begin to follow the user till



Fig. 6. A real situation of a user interacting with the robot

he/she gives the robot a waiting command, spreading the right hand horizontally and raising the left hand up. If the user tries to stop the robot following action and return to a ready status, he/she simply spreads the left hand horizontally and raises the right hand up. Figure 6 demonstrates the user standing in front of the robot at a proper work distance in an outdoor environment. In the right part of this figure, we also illustrate the user interface of the HMM-based hand gesture classifier installed on the robot.

6 Conclusions

In this paper, we have accomplished the recognition of pre-defined meaningful gestures consisting of continuous hand motions in an image sequence captured from a PTZ camera in real time, and an autonomous robot is directed by the meaningful gestures to complete some responses under an unconstraint environment. Four basic types of directive gestures made by a single hand have been defined such as moving forward, downward, leftward, and rightward. It results in twenty-four kinds of compound gestures from the combination of the directive gestures made by two hands. At present, we apply the most natural and simple way to select eight kinds of compound gestures employed in our developed human-robot interaction system, so that users can operate the robot effortlessly. From the experimental outcomes, the human-robot interaction system works by 7 frames per second on an average, and the resolution of captured images is 320×240 pixels using the PTZ camera. The average gesture recognition rate is more than 96%, which is quite satisfactory and encouraged to extend this system used in varied areas.

Acknowledgement. The authors thank the National Science Council of Taiwan for supporting this work in part under Grant NSC99-2221-E-011-132.

References

- [1] Soontranon, N., Aramith, S., Chalidabhongse, T.H.: Improved face and hand tracking for sign language recognition. In: Proc. of the Int. Conf. on Information Technology: Coding and Computing, Bangkok, Thailand, vol. 2, pp. 141–146 (2005)
- [2] Zhu, X., Yang, J., Waibel, A.: Segmenting hands of arbitrary color. In: Proc. of the IEEE Int. Conf. on Automatic Face & Gesture Recognition, Pittsburgh, Pennsylvania, pp. 446–453 (2000)
- [3] Mitra, S., Acharya, T.: Gesture recognition: A survey. *IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews* 37(3), 311–324 (2007)

- [4] Kim, K.K., Kwak, K.C., Chi, S.Y.: Gesture analysis for human-robot interaction. In: Proc. of the Int. Congress on Anti Cancer Treatment, Phoenix, Arizona, pp. 20–22 (2006)
- [5] Fahn, C.S., Chu, K.Y.: Real-time hand detection and tracking techniques for human-robot interaction. In: Proc. of the IADIS Interfaces & Human Computer Interaction Conference, Freiburg, Germany, pp. 179–186 (2010)
- [6] Liu, N., Lovell, B.C., Kootsookos, P.J., Davis, R.I.A.: Model structure selection and training algorithm for an HMM gesture recognition system. In: Proc. of the Int. Workshop in Frontiers of Handwriting Recognition, Brisbane, Australia, pp. 100–106 (2004)
- [7] Yoon, H., Soh, J., Bae, Y.J., Yang, H.S.: Hand gesture recognition using combined features of location, angle and velocity. Pattern Recognition 34(70), 1491–1501 (2001)
- [8] Korgh, A., Brown, M., Mian, I.S., Sjolander, K., Haussler, D.: Hidden Markov models in computational biology: Applications to protein modeling. Journal of Molecular Biology 235(5), 1501–1531 (1994)
- [9] Mohamed, M.A., Gader, P.D.: Handwritten word recognition using segmentation-free hidden Markov modeling and segmentation-based dynamic programming techniques. IEEE Transactions on Pattern Analysis & Machine Intelligence 18(5), 548–554 (1996)
- [10] Lawrence, R.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proc. of the IEEE 77(2), 257–286 (1989)

Manual and Accelerometer Analysis of Head Nodding Patterns in Goal-oriented Dialogues

Masashi Inoue^{1,2}, Toshio Irino³, Nobuhiro Furuyama^{4,5}, Ryoko Hanada⁶,
Takako Ichinomiya⁷, and Hiroyasu Massaki⁸

¹ Graduate School of Science and Engineering, Yamagata University, Yonezawa, Japan

² Collaborative Research Unit, National Institute of Informatics, Tokyo, Japan

³ Faculty of Systems Engineering, Wakayama University, Wakayama, Japan

⁴ Information and Society Research Division,

National Institute of Informatics, Tokyo, Japan

⁵ Department of Computational Intelligence and Systems Science,

Tokyo Institute of Technology, Tokyo, Japan

⁶ Graduate School of Clinical Psychology/Center for Clinical Psychology and Education,

Kyoto University of Education, Kyoto, Japan

⁷ Graduate School of Clinical Psychology, Kyoto University of Education, Kyoto, Japan

⁸ Department of Informatics, Sokendai, Tokyo, Japan

mi@yz.yamagata-u.ac.jp, irino@sys.wakayama-u.ac.jp,

furuyama@nii.ac.jp, {hanada,din95002}@kyoto-u.ac.jp,

ssaki@goo.jp

Abstract. We studied communication patterns in face-to-face dialogues between people for the purpose of identifying conversation features that can be exploited to improve human-computer interactions. We chose to study the psychological counseling setting as it provides good examples of task-oriented dialogues. The dialogues between two participants, therapist and client, were video recorded. The participants' head movements were measured by using head-mounted accelerometers. The relationship between the dialogue process and head nodding frequency was analyzed on the basis of manual annotations. The segments where nods of the two participants correlated were identified on the basis of the accelerometer data. Our analysis suggests that there are characteristic nodding patterns in different dialogue stages.

Keywords: Dialogue analysis, Face-to-face, Head nodding, Accelerometer.

1 Introduction

We shall describe our method of analyzing face-to-face interactions to find structural patterns in dialogues that may be useful in designing human-computer interactions. Note that various modalities have been investigated, such as the relationship between hand gestures and verbal miscommunications [1] and the relationship between body synchronicities and empathy in dialogues [2, 3].

Our focus is on head gestures, especially nodding. We quantitatively analyzed the usage of head nods in dialogues. Nodding is important because it is strongly tied with

the empathy that is the basis of the bond between participants of the interaction and has received much attention among various head movements [4]. The relationship between nods and the mood of the interaction can be better understood by observing task-oriented dialogues, especially those of psychological counseling. As Rogers pointed out [5], it is critically important for a therapist to listen to the client skillfully for the counseling to be successful. We expected that gestures, especially the time course of the nodding pattern, would play a certain role in giving an impression that the client is being attended by the therapist.

The two participants in counseling play quite different roles: therapists and clients. We measured the head nodding frequencies and their timings in relation to the changes in the dialogue's mood. Manual annotations made by a psychotherapist and a mechanical recording of head movements by using an accelerometer were used to analyze the relationship. The results showed that the changes in the nodding patterns are related to dialogue stages.

The present research project is concerned with the following two questions: 1) Does the frequency of nods of the therapist and the client change as therapeutic stages (initial stage, exploration stage, struggling stage, and closing stage) progress? 2) Does the phase-lag of nods between the therapist and the client change as the therapeutic stages progress?

1.1 Related Work

Head nodding in dialogues has been studied as a nonverbal behavior in the field of psychology [6, 7, 8]. However, these studies were either experimental ones that did not correspond to real dialogues or were summarizations of entire dialogues. We studied the time course of nodding patterns during actual task-oriented dialogues by using detailed manual coding and objective sensor data. The concept of dialogue stages is especially important when we want to apply our knowledge on effective interaction to computer systems that manage complex dialogue tasks rather than single-shot exchanges of messages. Another important aspect is that we contrasted dialogues conducted by specialists of different skill levels. Even psychotherapists can fail in counseling because the task is so complex. This aspect has not been well explored in the previous studies.

2 Data Collection

We recorded two counseling interviews. The therapists, clients, and topics varied. Each dataset was recorded on different dates and completed in a single interview session. The length of the interview was not pre-determined and could be controlled by the participants during the dialogue. The problems discussed were actual problems and not ones of role-plays that used specific roles and scenarios. The properties of the datasets are listed in Table 1. All participants used Japanese as their language of communication. The participants consented to being video recorded for research purposes. The first counseling was carried out by a novice therapist who is a graduate student studying psychotherapy. The second counseling was done by an expert therapist who supervises students majoring in psychotherapy. Both sessions completed

Table 1. Overview of datasets

Dataset	Duration (min.sec)	Therapist (level of experience)	Client
1	47.36	Female (beginner)	Male
2	29.49	Female (expert)	Female

**Fig. 1.** Microphone placed near the mouth**Fig. 2.** Accelerometers mounted in the occipital region

successfully but in different ways. In the first counseling, the client felt that he could organize his ideas and was relieved by talking. In the second counseling, the therapist proposed an intervention and the client agreed to try it with satisfaction.

All of the interviews were video-recorded. The participants faced each other, and the video camera was situated by their side and it captured images of their whole bodies. We used microphones (DPA 4060-BM) to clearly record their voices and triaxial head-mounted accelerometers (Kionix KXM52-1050) to record their head movements, as shown in Figure 1 and 2. Speech and acceleration were recorded at a 50-kHz (first data) and 20-kHz (second data) sampling rate. Although the accelerometers recorded triaxially, we used only the acceleration along the vertical axis; that is, we measured only the up and down movements that corresponded to nodding movements. The sensor itself was 1 square centimeter in size and could not be seen by the other person when mounted in the occipital region. Thanks to the non-invasive characteristics of our sensor mounting systems, the participants reported that they could maintain a natural dialogue.

3 Method

We conducted two analyses. The first was based on the manual coding of the differences in head nod frequencies at different stages of the dialogue. In the first analysis, a non-participating psychotherapist used ELAN annotation software to

manually annotate nods in both recordings. Each nod was identified as a segment from the beginning of the head movement to its end. The second analysis dealt with the degree of nodding synchronization between participants as revealed by the sensor time series data of the accelerometer. Therapists sometimes moved their head simultaneously with clients; while in other situations, they delayed in nodding. These differences may correspond to the task that the therapists have to conduct during the stages. The acceleration signals of head in the upward and downward directions were used as is regardless if they were manually annotated as nods or not. Cross correlation was calculated by using a 0.5 second window (sliding in 0.1 second increments) on the data from the two accelerometers. This analysis was meant to determine if therapists coordinate their nods with the clients' and if the degree of synchronization is associated with any events in the counseling sessions. The second analysis is explained in detail in the next section.

The usage of nods by therapists varies from stage-to-stage because the different stages correspond to the different roles taken to achieve therapeutic goals. In a past study, we found that there are different therapeutic stages in interview dialogues and they can be characterized by the occurrence patterns of speech types [9]. In this study, we tried to determine whether the nodding pattern changes according to the goals that therapists try to achieve in each stage.

3.1 Calculation of Cross Correlations

Acceleration data. The acceleration data were used as measures for the nodding movements. The accelerometer (KXM52-1050) outputs analog waveforms corresponding to the acceleration values along three orthogonal axes (X, Y, and Z). The acceleration waveforms and the speech of the therapist and the client were simultaneously digitized by using an eight channel data recorder (EZ7510) with a sampling rate of 50kHz (first data) and 20 kHz (second data). For the measurement setup, the nodding movements roughly corresponded to the Y values. The cross correlation was calculated for the Y data of the therapist and the client.

Cross correlation. The acceleration data was resampled at the sampling rate of 1 kHz and filtered by a notch filter of 60 Hz and a lowpass filter to reduce hum and high frequency noise. Denoting the denoised signal as $\mathbf{x} = (x_1, x_2, x_3, \dots, x_N)$, the normalized signal, \mathbf{x}_A , is

$$\mathbf{x}_A = \frac{\mathbf{x} - \bar{\mathbf{x}}}{\|\mathbf{x} - \bar{\mathbf{x}}\|}$$

where $\bar{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^N x_i$ (average) and $\|\mathbf{x}\| = \sqrt{\frac{1}{N} \sum_{i=1}^N x_i^2}$ (rms value). A window function, \mathbf{w} , was applied to the normalized signal, \mathbf{x}_A , to extract a short frame (or segment) of the data:

$$\mathbf{x}_w = \mathbf{x}_A \cdot \mathbf{w}.$$

The duration of the window function was 500 ms, and the frame shift was 100 ms. The window function had a flat part of 400 ms and tapers conforming a raised cosine

function (50 ms \times 2). The cross correlation $R_w(n_\tau)$ for one frame between the windowed signals for the client, \mathbf{x}_{w_1} , and the therapist, \mathbf{x}_{w_2} , is

$$R_w(n_\tau) = \sum_{n=-\infty}^{\infty} x_{w_1}(n + n_\tau) \cdot x_{w_2}(n),$$

where n_τ is the lag time between the client's nod and the therapist's nod in the sample. The correlation value, r , is the peak value (either positive or negative) of the cross correlation $R_w(n_\tau)$:

$$r = \max(\text{abs}(R_w(n_\tau))). \quad (1)$$

The correlation lag, n_{rp} , is defined as the lag at the peak. The correlation value was derived for every frame, and the sequence was derived for the whole therapy session. The maximum absolute value of r of all frames between the first and last three minutes of the session was used to normalize r . Thus, the peak value of r was unity in the session, excluding first and last three minutes.

4 Results and Discussion

4.1 Nodding Frequency

The nodding frequencies of the therapist (Th) and client (Cl) are shown in Figures 3 and 4, where the height of each bar represents the frequencies in 50 second segments and the horizontal axes represent time from the beginning of the session. To understand the overall trends rather than individual values in the time segments, we smoothed the frequencies by averaging the five preceding and five subsequent bins around the target bin and plotted these averages as lines.

The first analysis revealed that there were M-shaped changes in the therapists' nodding frequencies in both counseling dialogues, as shown in the upper graphs of Figure 3 and Figure 4. From the beginning of dialogue, the frequencies increased and then dropped for a while. Then it increased again, before decreasing toward the end of the session. The initial increase and final decrease can be understood as natural changes that occur at the starts and ends of dialogues. However, the reason for the drops in the middle of the dialogues is not clear. It is also interesting that the timing of the drops differed between the therapist and client. Accordingly, we closely looked at what was happening during these video segments. The lowest frequency of the first (beginner) therapist occurred when the client uttered "I'm not characterized as a parent" referring to the result of personality test he had taken. The drop might have happened because the client humbled himself by talking about his immaturity whereas the therapist wanted to mediate him and could not respond with an affirmative nod. The lowest frequency of the first client occurred when the therapist complimented the client by saying, "You can see yourself very well", but the client responded humbly by saying "No no, only this time". The drop seemingly occurred when the therapist was working on the client but the client could not accept her advice. These two drops were in different sequences but have a commonality that they happened when the

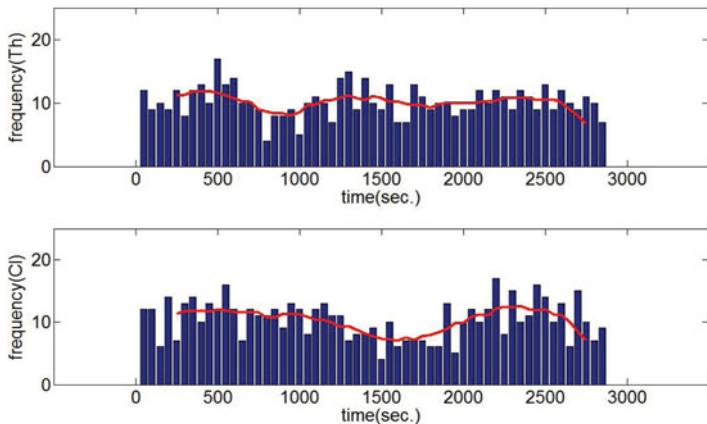


Fig. 3. Nodding frequency for the first dialogue

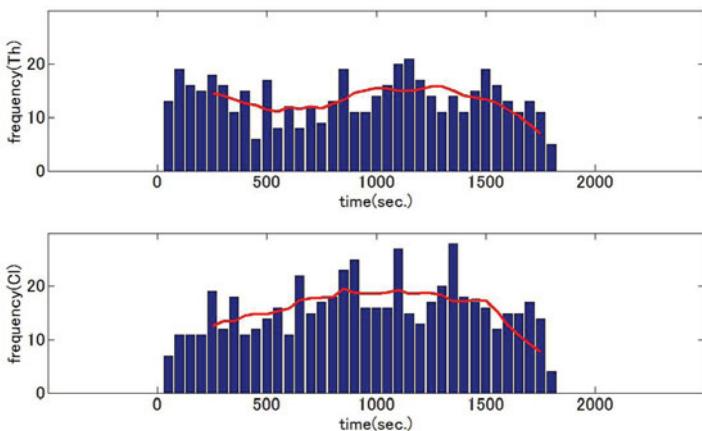


Fig. 4. Nodding frequency for the second dialogue

counseling went into the deepening stage of dealing with the client's problem. In that stage, utterances occurred to which the listeners would not or could not agree. That is, the nodding pattern changed when the counseling transited from the information gathering stage involving simple questions and answers to the stage where introspection played a central role and involved the client's monologue-like utterances.

The lower graph of Figure 3 shows that the nod frequency of the client interviewed by the novice therapist also had an M-shaped change. On the other hand, the client interviewed by the expert therapist showed an inverted U-shaped nodding frequency pattern (the lower graph of Figure 4). The lowest frequency of the first client occurred when the therapist was about to take the initiative in counseling. Until then, the client

had spontaneously detailed the problem and tried to figure out how to solve it, but she had failed in her mind to produce any clue to a solution. From that point, the therapist had to lead the session. This shift of dialogue stages, from the therapist listening to the client to her asking the client questions, might be reflected in the nodding frequency pattern of the therapist. On the other hand, there was no drop in frequency on the client's side, because the client did not go into the introspection but continued interaction with the therapist as their dialogue deepened. From the viewpoint of expertise, the first dialogue was led by the client and the inexperienced therapist simply reacted. In contrast, the second dialogue was coordinated by the experienced therapist, who was oriented toward solving the problem rather than letting the client analyze herself.

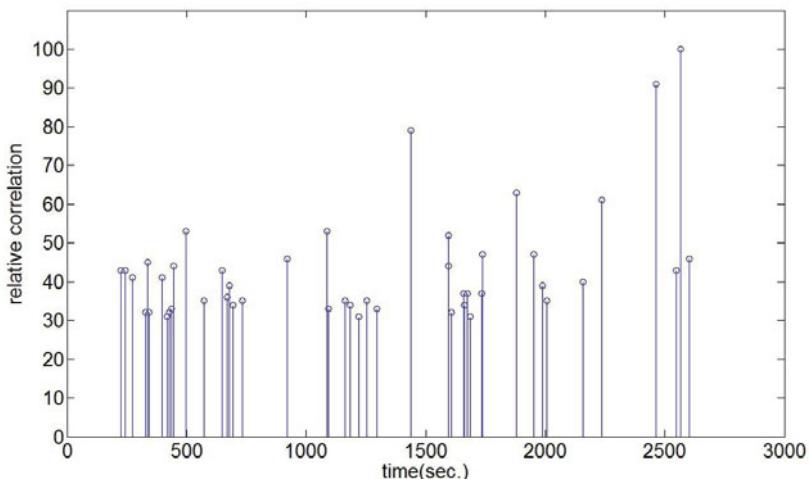


Fig. 5. Cross-correlation value of head accelerations for the first dialogue

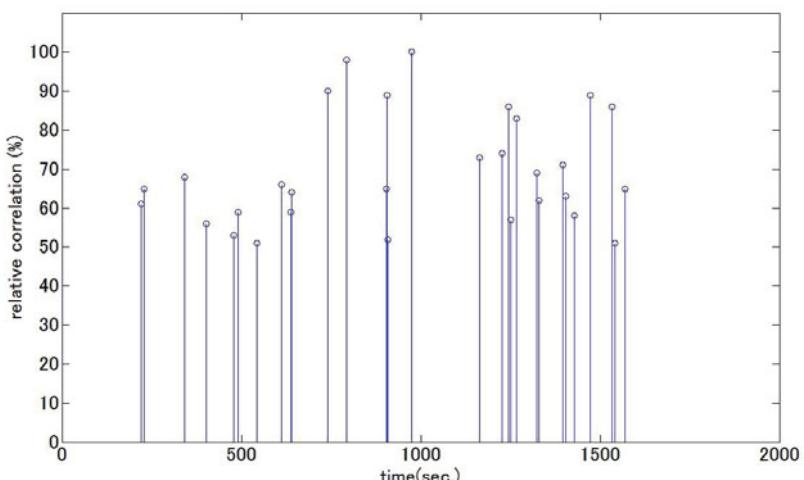


Fig. 6. Cross-correlation value of head accelerations for the second dialogue

4.2 Timing of Nodding

Figures 5 and 6 show the synchronies of head movements between therapist (Th) and client (Cl). The height of each bar represents the value of the cross correlation calculated by Equation 1, and the horizontal axes represent the time from the beginning of the session. We used the biggest acceleration value and speculated that the timing of the value may often correspond to when the head movement starts. For the ease of visual interpretation, we set threshold values as 30% and 50% of the maximum value in each figure and displayed bars higher than the thresholds.

In the first dialogue, the strongest synchrony occurred at the end of the dialogue when the client mentioned how nice his parents were. The therapist was trying to confirm the client's positive perception by using strongly correlated nodding. In the second dialogue, strong synchronies occurred in two places. In the first interaction, the client who had been replying negatively to the therapist eventually replied positively. At that point, the therapist seized upon the chance and tried to reach an agreement by nodding her head. In the second interaction, the therapist checked the facts surrounding the problem and asked for confirmation by the client in each utterance. The confirmation was accompanied by nodding. After this interaction, the client started talking lively.

The second analysis revealed that the degree of nodding synchronization was not segmental as seen in the frequency patterns; rather, it corresponded to particular mood shifts. By analyzing the content, we identified these interesting points where nodding synchronized differently between the two therapists. In the counseling of the novice therapist, nods synchronized as the client talked freely, while in the counseling of the experienced therapist, nods became strongly synchronized when the therapist started to intervene in the problem of the client. This difference can be interpreted as due to a similar difference of activity and passivity as in the first analysis.

5 Conclusion

To obtain insights for developing better human-computer interactions, we analyzed head nodding patterns in task-oriented dialogues of psychotherapy. By annotating nodding in videos of client-therapist interactions, we found that there are patterns of nodding frequency changes as the counseling proceeds. The degree of synchronization between two participants was measured by using accelerometers. We found that distinctive communication events occurred when head movements strongly correlated. In addition, by comparing therapists having different degrees of expertise, we identified the possibility that these head nodding patterns vary according to the skill level of the therapist.

In the future, we will examine the relationship between hand-annotated head nodding segments and acceleration signals and increase the amount of data to be analyzed in order to generalize the head nodding patterns found in this research. Other modalities such as speech and hand gestures will be also considered, especially as regards their interactions with nods.

Acknowledgments. This research was partially supported by Grants-in-Aid for Scientific Research 19530620 and 21500266, and the research grant from Kayamori Foundation of Informational Science Advancement.

References

1. Inoue, M., Ogihara, M., Hanada, R., Furuyama, N.: Gestural cue analysis in automated semantic miscommunication annotation. *Multimedia Tools and Applications* (to appear)
2. Maurer, R.E., Tindall, J.F.: Effect of postural congruence on client's perception of counselor empathy. *Journal of Counseling Psychology* 30, 158–163 (1983)
3. Nagaoka, C., Komori, M.: Body movement synchrony in psychotherapeutic counseling: A study using the video-based quantification method. *IEICE Trans. Info. Sys.* E91-D(6), 1634–1640 (2008)
4. Harrigan, J.A.: Proxemics, kinesics, and gaze. In: *The Handbook of Methods in Nonverbal Behavior Research*. Oxford University Press, New York (2008)
5. Rogers, C.: *Client-Centered Therapy: Its Current Practice, Implications, and Theory*. Houghton Mifflin, Boston (1951)
6. Matarazzo, J.D., Saslow, G., Wiens, A.N., Weitman, M., Allen, B.V.: Interviewer head nodding and interviewee speech durations. *Psychotherapy: Theory, Research & Practice* 1(2), 54–63 (1964)
7. McClave, E.Z.: Linguistic functions of head movements in the context of speech. *Journal of Pragmatics* 32, 855–878 (2000)
8. Maynard, S.: Interactional functions of a nonverbal sign: Head movement in Japanese dyadic casual conversation. *Journal of Pragmatics* 11, 589–606 (1987)
9. Inoue, M., Hanada, R., Furuyama, N.: Assessment of counseling stages to better understand the process toward solutions. In: *2008 Conference on Solution-Focused Practice*, Austin, Texas, November 12-16 (2008)

Facial Expression Recognition Using AAMICPF

Jun-Sung Lee, Chi-Min Oh, and Chil-Woo Lee

School of Electronics and Computer Engineering, Chonnam National University,
500-757 Gwang-Ju, Korea

{aliasim, sapeyes}@image.chonnam.ac.kr,
leecw@chonnam.ac.kr

Abstract. Recently, many interests have been focused on the facial expression recognition research because of its importance in many applications area. In the computer vision area, the object recognition and the state recognition are very important and critical. Variety of researches have been done and proposed but those are very difficult to solve. We propose, in this paper, to use Active Appearance Model (AAM) with Particle filter for facial expression recognition system. AAM is very sensitive about initial shape. So we improve accuracy using Particle filter which is defined by the initial state to particles. Our system recognizes the facial expression using each criteria expression vector. We find better result than using basic AAM and 10% improvement has been made with AAA-IC.

Keywords: Facial expression recognition, Active Appearance Model, Particle Filter.

1 Introduction

The face which is the best part on the human body represents inner psychological state except the language. Facial expression recognition is applied in many areas such as human computer interaction, multimedia information retrieval, medicine, and commercial product like digital camera and smart phone. So many research interests have been grown up in the facial expression analysis.

The processes for facial expression recognition are divided in two steps. The first, it extracts the features from the face expression. The main feature's areas are eyes, brows, nose, and mouth. The second, we design the recognizer for facial expression. Various techniques [4]-[8] have been proposed to detect and track facial features in face images. In general, two types of information are commonly utilized by these techniques. One is the image appearance of the facial features, which is referred as texture information; and the other is the spatial relationship among different facial features, which is referred as shape information.

In [7], a neural network is trained feature detector for each facial feature. Facial features will be located by searching the face image via the trained facial feature detectors. Similarly, Gabor wavelet networks are trained to locate the facial features in [3]. Since the shape information of the facial features is not modeled in both

techniques, they are prone to image noise. Therefore, in [8], a statistical shape model is built to capture the spatial relationships among facial features, and multi-scale and multi-orientation Gaussian derivative filters are employed to model the texture of each the texture of facial feature. However, only the shape information is used when comparing two possible feature point configurations, ignoring the local measurement for each facial feature. Such a method may not be robust in the presence of image noise. Cootes [5] proposed Active Appearance Model that uses the shape and texture model. Shape and texture model are applied by Principle Component Analysis (PCA) from hand-marked sample data. Then the feature can be extracted and tracked by regression model on facial image. This approach robust in terms of the noise and it has high accuracy but the initial state is very important for result. If the initial state is wrong, the result is almost failed.

To solve this problem, we proposed to use AAM with Particle Filter (PF) for finding good initial state. We define the initial shape to particle and generate the particle distribution around the center of the face. Then it performs AAM fitting on each particle and selects the result which has a minimal error.

After extracting the facial features, there are two approaches to recognize the facial expression. Those are static and temporal approaches. The static classifiers such as the neural networks, the support vector machines and the linear discriminant analysis attempt to recognize the facial expression using one image. The temporal classifiers such as the Hidden Markov Model and the recurrent neural networks attempt the facial expression recognition using a sequence of images. In this paper, we use the criteria expression model using a sequence of images. We choose the 3 facial expressions (happy, angry, and neutral).

2 Extract the Facial Features

In this section, we introduce the AAM and AAM with Particle Filter. Then, we select the feature for facial expression recognition.

2.1 Active Appearance Model (AAM)

In recent year there has been a growing interest in face modeling. Active Appearance Model [1],[4] are generative, parametric models that of a certain visual phenomenon that show both shape and appearance variations. These variations are represented by a linear model such as PCA. The face model can be made up from the training data using AAM. And the face features extracting is achieved by fitting the trained model to a input sequence.

AAM is represented by a triangulated mesh with one vertex. The shape vector scan expression as $s=(x_1, y_1, \dots, x_n, y_n)^T$ and shape variation is expressed by a linear combination of a mean shape s_0 and m shape basis vector s_i as

$$s = s_0 + \sum_{i=1}^m p_i s_i \quad (1)$$

Where p_i denote the i -th shape parameter, and $p=\{p_1, p_2, \dots, p_m\}$ is the shape parameter vector of AAM for input face image. The mean shape s_0 and m shape basis vectors s_i

are normally obtained by applying PCA to the training data, where the vertices of image are marked by hand. The i -th shape basis vector s_i is the i -th eigenvector that corresponds to the i -th largest Eigen value.

The training images are warped to the mean shape s_0 using the piece-wise affine warp that is defined between the corresponding triangles in the landmarked shape of the training images and the mean shape. Then, we can define the appearance as a shape normalized image $A(x)$ over the pixels x that belong to the inside of the s_0 . The appearance variation is expressed by a linear combination of a mean appearance $A_0(x)$ and n appearance basis vectors $A_i(x)$ as

$$A(x) = A_0(x) + \sum_{i=1}^n \alpha_i A_i(x), \quad (2)$$

where α denote the i -th appearance parameter, and $\alpha = \{\alpha_1, \alpha_2, \dots, \alpha_n\}$ is the appearance parameter vector of AAM for input face image. As with the shape model, the mean appearance $A_0(x)$ and n appearance basis vectors $A_i(x)$ are normally obtained by applying PCA to the training data. The i -th appearance basis vector (image) $A_i(x)$ is the i -th eigenvector that corresponds to the i -th largest Eigen value.

The Goal of AAM fitting is then to minimize the difference between the warped image and the appearance image. Thus, fitting AAM to a target image can be formulated as finding the model parameters of an AAM that minimize the following error as

$$E = \sum_{x \in s_0} [A(x) - A_0(x) - \sum_{i=1}^n \alpha_i A_i(x) - I(W(x; p, q))]^2, \quad (3)$$

where q is the global 2D pose parameter vector including the scale, rotation, and horizontal/vertical translation. A number of gradient-based fitting algorithms have been proposed, in which the Lucas-Kanade image matching algorithm [9], [10] is extended. In this paper, we selected the method that Ian Matthews and Simon Baker proposed Inverse compositional method.

2.2 AAM Based on Inverse Compositional Using Particle Filter (AAMICPF)

The fitting step of tradition AAMs used the single sample area. So if this area is wrong, all results are almost failed. To solve this problem, we use the Particle Filter which can be constructed multi-sample area.

Particle filter is a sequential Monte Carlo methodology where the basic idea is the recursive computation of relevant probability distributions using the concepts of importance sampling and approximation of probability distributions with discrete random measures. The fundamental idea of Particle filter approximates the filtered posterior (next) distribution (density) by a set of random particles (sampling) with associated weights. It weights particles based on a likelihood score and then propagates these particles according to a motion model. Particle filtering assumes a Markov Model for system state estimation. Markov model states that past and future states are conditionally independent of a given current state. Thus, observations are dependent only on current state. In this paper, we define the initial shape to particle and distribute particles according to particle parameters (numbers, range, scale and rotation of particles). We perform AAM fitting algorithm in each particle location and the final result accomplished with the particle which has a minimal error.

2.3 Selecting Features for Facial Expression Recognition

We use the 68 vertices for constructing the shape model (Brows 12, eyes 10, nose 12, mouth 19, jaw 15). When the facial expression is changed, the several features change. Those parts are brows, eyes and mouth. So we just use the 42 vertex except nose and jaw vertex.

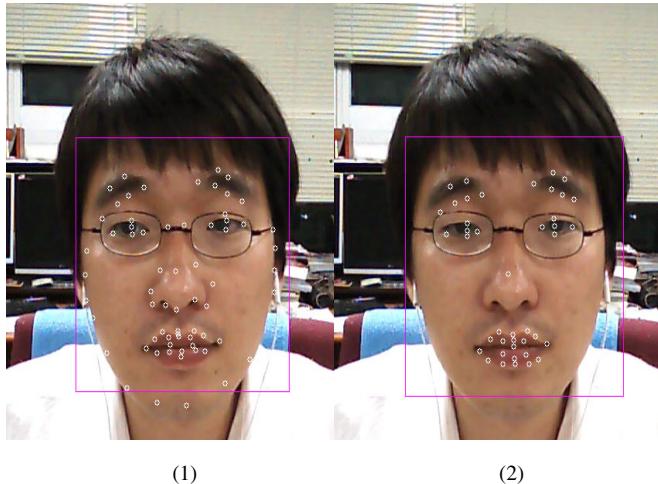


Fig. 1. Result of extracting features (1) AAMICPF (2) feature for facial expression

3 Facial Expression Recognition

In this paper we recognize three facial expressions (happy, angry, and neutral). And we use the criteria vector for each facial expression. This method divides the two steps: training and recognition step. In training step, we extract the features using proposed AAMICPF and select the feature for facial expression. So we call this feature set F and denote the equation 4. We calculate the center of the feature and alignment and normalize using the equation 5, 6 and 7.

$$f = \{f_1, f_2, \dots, f_i\} (1 \leq i \leq 42) \quad (4)$$

$$f_M = \frac{\sum_{i=1}^{42} f_i}{42} \quad (5)$$

$$F' = F - f_M \quad (6)$$

$$\bar{F} = \frac{F'}{|F'|} \quad (7)$$

We calculate normalized distance histogram using iterating this step. Then, we normalize histogram of each expression using equation 8. This will be criteria vector.

$$\overline{F_M} = \frac{1}{N} \sum_i^N \overline{F_i} \quad (8)$$

Similarity, in recognition step, we extract the features using AAMICPF from input image and we calculate F_I using equation 4, 5, 6, 7 and 8. Then we calculate MSE using equation 9 between each expression criteria vector F_M and F_I .

$$MSE(F_I) = \sqrt{(\overline{F_I} - \overline{F_M})^2} \quad (9)$$

The final result is the minimum MSE value for facial expression recognition. Figure 2 shows the main working block diagram of facial expression recognition with AAMICPF in simplified form.

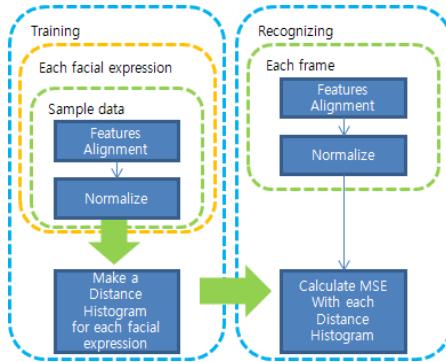


Fig. 2. Block diagram of facial expression recognition with AAMICPF algorithm

4 Experiments and Results

The proposed facial expression recognition technique is implemented using C++ on a PC with a i3 2.93GHz CPU and a 2GB RAM. The resolution of the captured images is 640x480 pixels, and the built system runs at approximately 25 fps.

We collected the database which included 10 Korean peoples (8 male and 2 female whose ages are between 24 and 30, and 5 peoples of them wear the glasses) and had three expressions (happy, angry and neutral) for testing and training. It consists of 150 frames for each expression and total 4500 frames. We use 50% of the database for training, the others is for testing. The Table 1 show that compare performance with traditional AAM, AAMIC and AAMICPF using facial expression recognition.

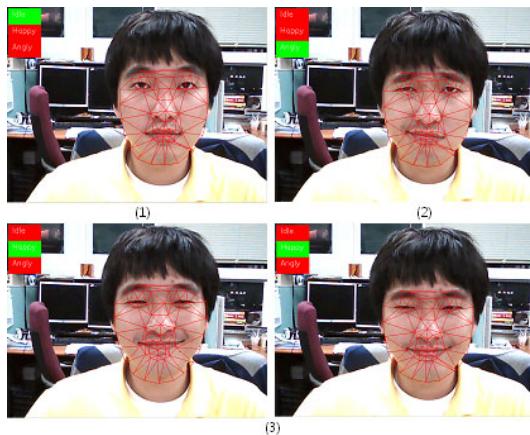


Fig. 3. Result of facial recognition (1) neutral (2) angry (3) happy

Table 1. Performance assessment with other algorithm (%)

	Neutral	Happy	Angry	Average
AAMICPF	82.12	92.40	93.65	89.39
AAMIC	70.38	79.60	80.25	76.74
Trad. AAM	68.23	75.00	76.43	73.22

5 Conclusion

In this paper, we propose to solve initial shape location problem using particle filter. Our System clearly improves the accuracy, and performs the facial expression recognition with proposed system and compares the result with traditional AAM and AAMIC.

In future work, we will improve more about our algorithm for extracting good features. Besides, our future research will concern more about the improvement of facial expression recognition and considering application of extracting facial features.

Acknowledgments. This research is supported by Ministry of Culture, Sports and Tourism (MCST) and Korea Culture Content Agency(KOCCA) in the Culture Technology(CT) Research & Development Program 2010.

References

1. Matthews, I., Baker, S.: Active appearance models revisited. International Journal of Computer Vision 60(2), 135–164 (2004)
2. Cohen, I., Sebe, N., Chen, L., Garg, A., Huang, T.: Facial expression recognition from video sequences: temporal and static modeling. Computer Vision and Image Understanding 91(1), 160–187 (2003)

3. Toyama, K., Feris, R.S., Gemmell, J., Kruger, V.: Hierarchical wavelet networks for facial feature localization. In: International Conference on AFGR (2002)
4. Edwards, G., Taylor, C., Cootes, T.: Active appearance models. IEEE Transaction on Pattern Analysis and Machine Intelligence 23(5), 681–685 (2001)
5. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. In: Burkhardt, H., Neumann, B. (eds.) ECCV 1998. LNCS, vol. 1407, pp. 484–498. Springer, Heidelberg (1998)
6. Wiskott, L., Fellous, J.M., Kruger, N., Malsburg, C.V.: Face recognition by elastic graph matching. IEEE Transactions on PAMI 19(7) (1997)
7. Reinders, M.J., Koch, R.W., Gerbrands, J.: Locating facial features in image sequences using neural networks. In: International Conference on AFGR (1996)
8. Burl, M., Leung, T., Perona, P.: Face localization via shape statistics. In: International Conference on AFGR (1995)
9. Baker, S., Matthews, I.: Lucas-Kanade 20 years on: a unifying framework: Part I. CMU Technical Report CMU-RI-TR-02-16 (2002)
10. Gross, R., Matthews, I., Baker, S.: Lucas-Kanade 20 years on: a unifying framework: Part 3. CMU Technical Report CMU-RI-TR-03-05 (2003)

Verification of Two Models of Ballistic Movements

Jui-Feng Lin¹ and Colin G. Drury²

¹ Department of Industrial Engineering & Management,
Yuan Ze University, Chung-Li, Taiwan 320

² Department of Industrial & Systems Engineering,
State University of New York at Buffalo, Buffalo, New York 14260, USA

Abstract. The study of ballistic movement time and ballistic movement variability can help us understand how our motor system works and further predict the relationship of speed-accuracy tradeoffs while performing complex hand-control movements. The purposes of this study were (1) to develop an experiment for measuring ballistic movement time and variability and (2) to utilize the measured data to test the application of the two models for predicting the two types of ballistic movement data. In this preliminary study, four participants conducted ballistic movements of specific amplitudes, by using a personal computer, a drawing tablet and a self-developed experimental program. The results showed that (1) the experiment successfully measured ballistic movement time and two types of ballistic movement variability, (2) the two models described well the measured data, and (3) a modified model was proposed to fit better the variable error in the direction of the movement.

Keywords: Fitts' law, aiming movement, ballistic movement, hand-control movement, movement time, and end-point variability.

1 Introduction

Hand-control movements occur in numerous everyday activities, such as approaching and grasping an object or pointing a computer cursor to move to an icon. The speed-accuracy tradeoff relationship for hand-control movements has been one of the most important issues studied by researchers in this field. Faster movements are usually spatially less accurate, and inversely, movements desiring greater spatial accuracy are performed at the expense of slower speeds.

Since Woodworth's [1] pioneer study, researchers have studied the speed-accuracy tradeoff in normal hand-control movements. The speed-accuracy tradeoffs are commonly studied for two types of movements: self-paced aiming movements and self-paced tracking movements. Fitts' [2] law and Drury's [3] model have been validated to predict well the two types of tradeoffs, respectively.

Beyond experimental modeling of the tradeoffs, researchers further studied how the two types of tradeoff result from our motor control mechanism. Because the utility of visual feedback on movement accuracy is intermittent, instead of continuous, a hand-control movement is typically composed of one or more than one sequential sub-movement. These sub-movements are programmed by the brain based only on one-time visual feedback, and are thus called "ballistic movements". In fact, these

ballistic movements play an important role in determining movement speed and movement accuracy in a variety of complex motor-control tasks [4-6].

With a comprehensive understanding of ballistic movements, individuals' performance in performing complex movements in terms of speed and accuracy can be predicted based on their measured ballistic movement properties, without experimental measurements of these complex tasks. To achieve such goals, the main research questions related to ballistic movements are: (1) how a hand-control movement is composed of ballistic movements, (2) the time required to perform ballistic movements, and (3) the endpoint variability of ballistic movements [7].

1.1 Ballistic Movement

The definitions of ballistic movement are different from study to study in the literature. As Gan & Hoffmann [8] noted, some psychologists referred a ballistic movement as a rapid movement where there is no visual feedback [e.g., 9, 10-12], while some neurophysiologists classified it as a rapid movement in the absence of corrective feedback of exteroceptive origin such as visual and auditory [e.g., 13, 14-17]. In practical terms, although with a valid mathematical model, Hoffmann [18] and Gan & Hoffmann [8] determined that aiming movements with index of difficulty [see 2] equal to or less than three are ballistic.

To avoid confusion, in this article, specific definitions of the term, "ballistic movement", were made based on Craik's [19] definitions and with an additional one. The three definitions are: (1) a ballistic movement is a unit movement executed with a movement order programmed by the brain, (2) it executes as a whole and would not be subject to interference or modification until its completion, and (3) a ballistic movement is a rapid movement performed voluntarily without temporal constraint (such as paced by a metronome).

To understand our motor control mechanism, two properties of ballistic movements should be studied – ballistic movement time and ballistic movement accuracy, typically measured by its end-point variability. Relevant studies of these two properties are next reviewed.

1.2 Ballistic Movement Time

Ballistic movement time represents the time required for performing a ballistic movement. Although Fitts' [2] law is commonly accepted by researchers to modeling aiming movements, Gan & Hoffmann [8] argued that Fitts' law was inadequate for describing aiming movements with a small index of difficulty (i.e., $ID \leq 3$). They stated that these movements were performed ballistically – visual feedback might not be helpful once the movements were executed. Hence, they proposed and validated a better model, Equation 1, for modeling the relationship between the movement time and movement amplitude.

$$t_{ballistic} = a + b\sqrt{d_u} \quad (1)$$

where $t_{ballistic}$ is ballistic movement time, a and b are experimentally determined constants, and d_u is uncontrolled movement distance, usually referred to as the amplitude.

1.3 Ballistic Movement Variability

Ballistic movement variability describes the endpoint variability of a ballistic movement. Because of noise existing in our motor control mechanism, the ultimate endpoint of a ballistic movement may not be exactly at the aimed point [20, 21]. While performing the movement repeatedly, the probability of its endpoint location would be a bivariate normal distribution around the aimed point [1, 22].

Beggs, Howarth and their colleges [23-26] conducted a series of studies that were related to modeling ballistic movement variability. To study how visual information affects movement accuracy, Beggs & Howarth [23] conducted an experiment in which their participants moved a pencil from a home position near their shoulders to hit a target attached on a vertical plate placed in front of them. The movement distance was about 50 cm and the movement speed was paced by a metronome with different frequencies. To ensure that the movements were uncontrolled (i.e., moving without visual feedback), the room illumination was turned off before reaching the vertical plate at several distances from the plate. For the uncontrolled part of these movements, Beggs & Howarth [26] found that Equation 2, developed by Howarth et al. [25], could well predict that the standard deviation of the aiming error (σ) was proportional to the uncontrolled distance (d_u).

$$\sigma^2 = c + d \times d_u^2 \quad (2)$$

where c and d are experimentally determined constants, and d_u is uncontrolled movement distance, as in Equation 1.

1.4 Research Objectives

At first glance, the movements modeled by Gan & Hoffmann [8] and Beggs & Howarth [26] were ballistic and hence Equation 1 and Equation 2 seem that they can directly be utilized to predict the two properties of ballistic movement. On closer consideration, further investigation is required to validate their applications. First, the movements studied by Gan & Hoffmann [8] could not represent the population of all ballistic movements as defined in this study. Because, the ballistic movements studied by Gan & Hoffmann [8] were aiming movements with small ID . The authors stated that these movements were different from Fitts-type aiming movements because these movements were all performed under 200 milliseconds during which movement accuracy would not be improved by visual feedback during the execution. Nevertheless, the ballistic movements as defined in this study were not restricted to certain movement time. Secondly, the movements studied by Beggs & Howarth [26] did not exactly qualify under our definitions for a ballistic movement. The ballistic part of movements modeled by Beggs & Howarth [26] were performed with nonzero starting speeds as the room illumination was turned off. Also, the aiming movements studied by Beggs & Howarth [26] were paced movements; paced by a metronome. It has been shown that aiming movement accuracy would be diminished if the movements are paced [21]. Finally, the aiming error reported in Beggs & Howarth [26] was only measured perpendicular to the movement direction.

The purpose of the current pilot study was to test an experiment for measuring the ballistic movement time and both the perpendicular and parallel ballistic movement variability. Measured ballistic movement properties were, naturally, utilized to test the application of Equation 1 and Equation 2.

2 Method

2.1 Participants and Apparatus

Two male and two female graduate students, aged from 25-30 years, were recruited to participate in this pilot study. They were all right-handed with normal or corrected-to-normal vision.

A personal computer (PC) with a 17" (432 mm) LCD monitor of resolution 1280×1024 pixels resolution was used. The PC ran Visual Basic using a self-designed experimental program that both displayed the experimental tasks and measured task performance. An Intous 3 305 mm \times 488 mm drawing tablet with a tablet stylus was utilized as the input device to perform ballistic movements. The movement distance ratio between the tablet and the computer screen was set as 1:1, equalizing visual and physical movement distances between the screen and the tablet.

2.2 Experimental Setting and Procedures

While conducting the experiment, the participants sat alongside a dual surface adjustable table on which the monitor and the tablet were placed on the rear and the front surfaces, respectively. Both the monitor and the tablet were adjusted to heights where the individual participants felt comfortable. To keep the friction between moving hand and the tablet surface small and constant, participants wore a nylon half-finger glove and kept resting their hands on the tablet surface. A cardboard screen was placed between their eyes and the tablet to hide the visual feedback from their moving hands so that the only visual feedback was from the monitor screen.

To perform the experimental task, the participants drew a line from a start point to the center of a cross target with a manipulated distance. The movements were all performed from left to right parallel to the coronal plane. The tasks started by pressing down the stylus cursor on the start point and then moving toward the cross target. Once the cursor was moved away from the start point, the cursor and the cross target disappeared and the movement time started to record. When the movement stopped, the information about the cross target and the endpoint of that movement were immediately displayed on the screen. By clicking the cursor on any location of the screen, the participants could continue on the next trail.

2.3 Experimental Variables

Independent variables of the experiment were: ballistic movement distance, X shift and Y shift of start-point location. The 14 values of ballistic movement distance (d_u) were 8, 17, 32, 53, 80, 113, 152, 197, 248, 305, 368, 437, 512, & 593 pixels (1 pixel $\equiv 0.266$ mm). Two X shifts (+100, -100 pixels) and two Y shifts (+40, -40 pixels)

were manipulated to generate four different start-point locations, helping eliminate learned kinesthetic feedback. Every experimental combination was replicated five times, resulting in a total of 280 trials. All the trials were randomly conducted by each participant, taking about half an hour to finish. There was a half-hour practice before the formal measurement.

Three dependent variables, comprising ballistic movement time ($t_{ballitic}$), longitudinal error and lateral error, were automatically recorded by the experimental program after every completed experimental trial. The longitudinal error, or X-error, was the longitudinal discrepancy between a ballistic movement endpoint and the target cross. Similarly, the lateral error, or Y-error, was the perpendicular discrepancy between the endpoint and the target cross.

3 Results

3.1 Ballistic Movement Time

Analysis of variance was performed on the movement times, using a mixed model with Distance, X Shift and Y Shifts as fixed effects and Participant as random, in which the two-way, three-way, and four-way interaction effects among all the effects were analyzed. The results showed significant main effects of Participant ($F_{3,896} = 12.42$, $p < 0.001$) and Distance ($F_{13,896} = 64.12$, $p < 0.001$) and two-way interaction effects of Participant \times Distance ($F_{3,896} = 7.55$, $p < 0.001$). These significant effects showed that (1) participants performed the movements with different ballistic movement times, (2) the increase of ballistic movement distance results in increased ballistic movement time, and (3) these rates of increase of ballistic movement time were different for the participants, representing the interaction effect of Participant \times Distance.

Because a significant main effect of Distance was found, the application of Gan & Hoffmann's [8] model could be tested. The means of ballistic movement time ($t_{ballitic}$) were regressed on to the square root of ballistic movement distance ($\sqrt{d_u}$) to give the slopes and intercepts. The model fitted the data very well. It accounted for 98.1 % variance of the overall participants' data and at least 95.2 % variance of individual participants' data. The regression lines of the overall participants and of individual participants are shown in Fig. 1, which also shows good model fittings.

3.2 Ballistic Movement Variability

End-point errors were measured as both longitudinal error (X error) and lateral error (Y error) relative to the movement direction. Both were verified to be normally distributed ($p < 0.001$). The errors consisted of constant error and variable error. To analyze whether the independent variables had significant effects on these two types of errors, five replications of each experimental combination were utilized to calculate the constant error and the variable error (measured by variance). However, only the results of variable error are discussed in this article.

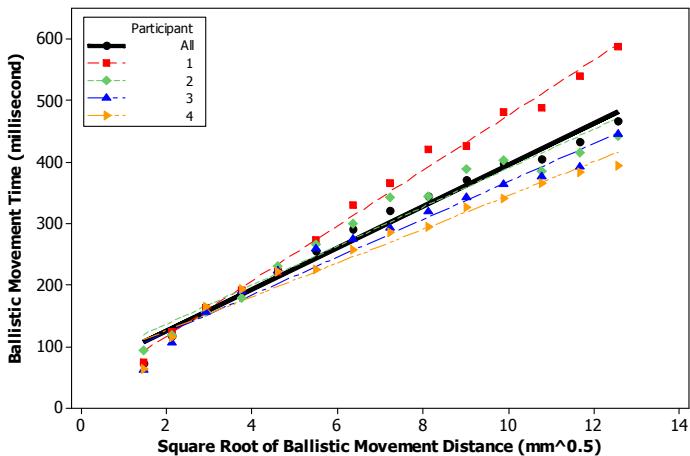


Fig. 1. The relationship between ballistic movement time ($t_{ballistic}$) and the square root of distance ($\sqrt{d_u}$)

Analysis of variance was performed on X-variable error and Y-variable error, using a mixed model with Distances, X Shift and Y Shift as fixed effects and Participant as random, with all interaction effects included. For X-variable error, there were significant effects of Participant ($F_{3,126} = 6.23, p = 0.001$) and Distance ($F_{13,126} = 13.77, p < 0.001$). The significant effect of Participant showed that the participants performed the movements with different variability. The significant effect of Distance showed that the increase of ballistic movement distance resulted in an increased X-variable error. In terms of Y-variable error, the only significant effect found was Distance ($F_{13,126} = 14.75, p < 0.001$), where an increase in ballistic movement distance also resulted in an increased Y-variable error.

With Distance having significant effects on both the X-variable error (σ_x^2) and Y-variable error (σ_y^2), Equation 2 could be tested for each error direction. The two error variances, calculated from the raw data for each distance, were regressed on to d_u^2 to give the slopes, intercepts, and r^2 values. For X-variable error, Equation 2 accounted for 92.4 % variance of the overall participants' data and at least 78.8 % variance of individual participants' data. For Y-variable error, Equation 2 accounted for 97.8 % variance of the overall participants' data and at least 79.9 % variance of individual participants' data. The regression lines of the overall participants' σ_x^2 and σ_y^2 data are shown in Fig. 2, where X-variable error is about four times larger than Y-variable error.

While d_u^2 was treated as the predictor in the linear regression model, the model did not fit the X-variance error as well as the Y-variance error. As shown in Fig. 2, the regression line overestimated the X-variance error for shorter distances and longer distances, but underestimated the error for middle distances. Thus, a different model (Equation 3) in which distance (d_u) was the predictor was fitted to the data.

$$\sigma^2 = e + f \times d_u \quad (3)$$

where e and f are experimentally determined constants, and d_u is again the uncontrolled movement distance.

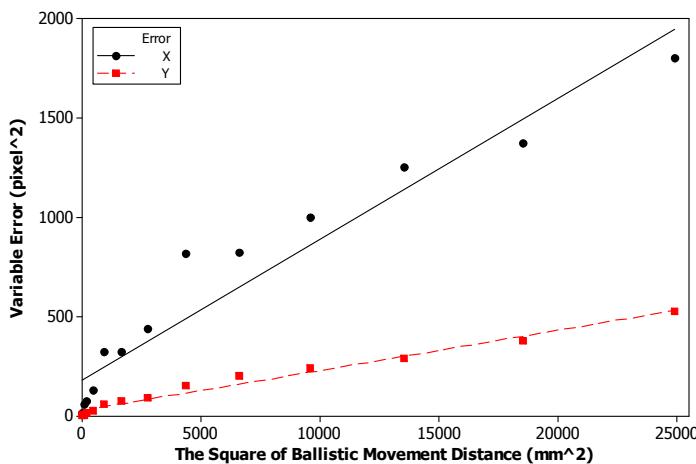


Fig. 2. Relationships between two types of variable errors (σ_X^2 and σ_Y^2) and the square of distance (d_u^2)

For X-variable error, Equation 3 accounted for 98.5 % variance of the overall participants' data and at least 81.9 % variance of individual participants' data. For Y-variable error, Equation 3 accounted for 96.5 % variance of the overall participants' data and at least 70.7 % variance of individual participants' data. Further, the regression lines of the overall participants' σ_X^2 and σ_Y^2 data are shown in Fig. 3. As shown in the figure, Equation 3 can fit the X-variable error somewhat better than Equation 2.

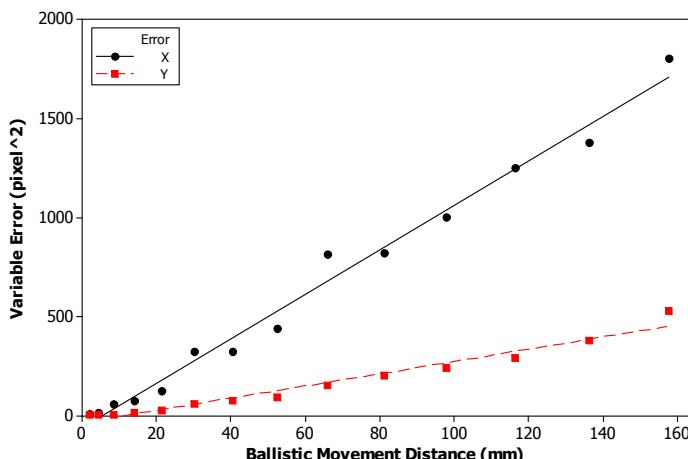


Fig. 3. Relationships between two types of variable errors (σ_X^2 and σ_Y^2) and distance (d_u)

4 Discussion

This study successfully measured ballistic movement time and ballistic movement variability, both perpendicular and parallel to movement direction when participants used a drawing tablet. The ballistic movement time ranged widely, from 50 milliseconds to 600 milliseconds presumably because of the wide range of d_u values used as well as the tablet/screen set-up. Furthermore, we found X-variable error (σ_x^2) was about four times larger than Y-variable error (σ_y^2).

The testing of Gan & Hoffmann's [8] model (i.e., Equation 1) showed that the model predicted the relationship between ballistic movement time and the square root of ballistic movement distance very well, despite the wider range of d_u values used. Gan & Hoffmann [8] used d_u values from 40 mm to 250 mm while we extended this range at the lower end, ranging from 2.1 mm to 157 mm. In Gan & Hoffmann [8], Equation 1 was tested with aiming movements with small IDs and with short movement time (less than 200 milliseconds). In this study, we further verified the prediction of Equation 1 for the ballistic movements that were performed even as long as 600 milliseconds.

The testing of Howarth et al.'s [25] model (i.e., Equation 2) showed that the model also predicted well the relationship between two types of ballistic movement variability and the square of ballistic movement distance. Again, the range of values of d_u values used here was different: 2-157 mm rather than Howarth et al.'s [25] 50-450 mm. However, it predicted Y-variable error better than X-variable error. To better fit the data, we tested a modified model (Equation 3), which utilized distance, instead of the squared distance, as the predictor. The comparison of Equation 2's and Equation 3's fitting results showed that Equation 2 predicted the Y-variable error better than Equation 3 ($r^2 = 0.978$ vs. 0.965), but Equation 3 predicted X-variable error better than Equation 2 ($r^2 = 0.985$ vs. 0.924). The prediction difference of X-variable error between Equation 2 and Equation 3 can be clearly seen while comparing Fig. 2 and Fig. 3. It seemed that movement variability measure perpendicular and parallel to movement direction possess different features. Since Equation 2 was theoretically developed based on Y-variable error [25], it might not valid for predicting X-variable error. Note that both formulations produced very good fits to the data set.

5 Conclusions

This study tested Gan & Hoffmann's [8] model and Howarth et al.'s [25] model for predicting movement time and two types of end-point variability while performing ballistic movements. The experiment successfully measured ballistic movement data by drawing tablet for response and a PC monitor for feedback. Gan & Hoffmann's [8] model predicted well the ballistic movement time even when the time were extended to 600 milliseconds. Also, Howarth et al.'s [25] model predicted well the ballistic movement variability. However, we found that X-variable error was fitted better by Howarth et al.'s [25] model and Y-variable error was fitted better by a modified model (Equation 3). In terms of variance error, X-variable error was about four times

larger than Y-variable error. Larger numbers of participants should be tested in future research to study individual differences, and Equation 3 should be theoretically validated.

Acknowledgments. We would like to acknowledge the Mark Diamond Research Fund of the Graduate Student Association at the University at Buffalo, The State University of New York for funding the drawing tablet and partial subject participation fees. Also, we would like to acknowledge the grant support from Taiwan National Science Council (NSC 99-2221-E-155 -066) for funding the paper submission and presentation.

References

1. Woodworth, R.S.: The accuracy of voluntary movement. *The Psychological Review Monographs Supplement* 3(13), 1–114 (1899)
2. Fitts, P.M.: The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology* 47, 381–391 (1954)
3. Drury, C.G.: Movements with lateral constraint. *Ergonomics* 14(2), 293–305 (1971)
4. Keele, S.W.: Movement control in skilled motor performance. *Psychological Bulletin* 70(6), 387–403 (1968)
5. Drury, C.G., Montazer, M.A., Karwan, M.H.: Self-paced path control as an optimization task. *Transactions on Systems, Man, and Cybernetics* 17(3), 455–463 (1987)
6. Montazer, M.A., Drury, C.G., Karwan, M.H.: An optimization model for self-paced tracking on circular courses. *Transactions on Systems, Man, and Cybernetics* 18(6), 908–915 (1988)
7. Lin, J.-F., Drury, C., Karwan, M., Paquet, V.: A general model that accounts for Fitts' law and Drury's model. In: Proceedings of the 17th Congress of the International Ergonomics Association, Beijing, China, August 9–14 (2009)
8. Gan, K.-C., Hoffmann, E.R.: Geometrical conditions for ballistic and visually controlled movements. *Ergonomics* 31, 829–839 (1988)
9. Welford, A.T.: *Fundamentals of skill*. Barnes & Noble, New York (1968)
10. Schmidt, R.A., White, J.L.: Evidence for an error detection mechanism in motor skills: A test of Adam's closed-loop theory. *Journal of Motor Behavior* 4(3), 143–153 (1972)
11. Glencross, D.J.: Control of skilled movement. *Psychological Bulletin* 84(1), 14–29 (1977)
12. Brooks, V.B.: Motor-program revisited. In: Talbott, R.E., Humphrey, D.R. (eds.) *Posture and Movement*. Raven Press, New York (1979)
13. Angel, R.W.: Electromyography during voluntary movement: The two-burst pattern. *Electroencephalography and Clinical Neurophysiology* 36, 493–498 (1974)
14. Hallett, M., Marsden, C.D.: Effect of perturbations on the EMG pattern of ballistic movements in man. *Clinical Neurophysiology* 43, 596 (1977)
15. Hallett, M., Shahani, B.T., Young, R.R.: EMG analysis of stereotyped voluntary movements in man. *Journal of Neurology, Neurosurgery and Psychiatry* 38, 1154–1162 (1975)
16. Water, P., Strick, P.L.: Influence of 'strategy' on muscle activity during ballistic movements. *Brain Research* 207, 189–194 (1981)
17. Denier Van Der Gon, J.J., Wadman, W.J.: Control of fast ballistic human arm movement. *Journal of Physiology* 271, 28–29 (1977)

18. Hoffmann, E.R.: An ergonomics approach to predetermined motion time systems. In: Proceedings from the 9th National Conference, Institute of Industrial Engineers, Australia, pp. 30–47 (1981)
19. Craik, K.J.W.: Theory of the human operator in control systems I: The operator as an engineering system. *British Journal of Psychology* 38, 56–61, 142–148 (1947)
20. Meyer, D.E., Smith, J.E.K., Wright, C.E.: Models for the speed and accuracy of aimed movements. *Psychological Review* 89(5), 449–482 (1982)
21. Schmidt, R.A., Zelaznik, H.N., Hawkins, B., Frank, J.S., Quinn, J.T.: Motor-output variability: A theory for the accuracy of rapid motor acts. *Psychological Review* 86, 415–451 (1979)
22. Crossman, E.R.F.W., Goodeve, P.J.: Feedback control of hand-movement and Fitts' law. *Quarterly Journal of Experimental Psychology* 35A, 251–278 (1963/1983)
23. Beggs, W.D.A., Andrew, J.A., Baker, M.L., Dove, S.R., Fairclough, I., Howarth, C.I.: The accuracy of non-visual aiming. *Quarterly Journal of Experimental Psychology* 24, 515–523 (1972)
24. Beggs, W.D.A., Sakstein, R., Howarth, C.I.: The generality of a theory of the intermittent control of accurate movements. *Ergonomics* 17(6), 757–768 (1974)
25. Howarth, C.I., Beggs, W.D.A., Bowden, J.M.: The relationship between speed and accuracy of movement aimed at a target. *Acta Psychologica* 35, 207–218 (1971)
26. Beggs, W.D.A., Howarth, C.I.: The accuracy of aiming at a target - some further evidence for intermittent control. *Acta Psychologica* 36, 171–177 (1972)

Gesture Based Automating Household Appliances

Wei Lun Ng, Chee Kyun Ng, Nor Kamariah Noordin, and Borhanuddin Mohd. Ali

Department of Computer and Communication Systems Engineering,

Faculty of Engineering, University Putra Malaysia,

UPM Serdang, 43400 Selangor, Malaysia

william_2909@hotmail.com,

{mpnck, nknordin, borhan}@eng.upm.edu.my

Abstract. Smart homes can be a potential application which provides unobtrusive support for the elderly or disabled that promote independent living. In providing ubiquitous service, specially designed controller is needed. In this paper, a simple gesture based automating controller for various household appliances that includes simple lightings to complex electronic devices is introduced. The system uses the gesture-based recognition system to read messages from the signer and sends command to respective appliances through the household appliances sensing system. A simple server has been constructed to perform simple deterministic algorithm on the received messages to execute matching exercise which in turn triggers specific events. The proposed system offers a new and novel approach in smart home controller system by utilizing gesture as a remote controller. The adapted method of this innovative approach, allows user to flexibly and conveniently control multiple household appliances with simple gestures.

Keywords: Gesture, smart home, stand-alone server, flex sensor, deterministic algorithm, remote controller.

1 Introduction

Generally, smart home (or commonly known as home automation and domotics) is a home equipped with various sensors, actuators and other technology that assists its resident in performing daily activities [1]. Assistance is done using various integrated hardware devices and software application which provide ubiquitous services. As computer technology advances, current smart home systems have transited from centralized one-system architecture to loosely-coupled distributed one. It is composed of cooperative agent that integrates heterogeneous ubiquitous computing in providing services [2].

In providing ubiquitous service, smart home uses various types of controllers to control household appliances. The smart home controller is the most important element in a smart home environment as it provides both controls and gateway between users and household appliances [3]. Occasionally, the gateway are connected to the home network; composed of communications with household appliances such as broadband modem, router, PCs, wireless access point, entertainment peripherals,

and other electronic devices. It allows users to control household appliances remotely through the external network like Internet or mobile network.

This paper presents a new approach in controlling household appliances by using gesture based controller. A gesture based recognition system reads the gestures produced by the signer. The controller is able to control from simple to complex household appliances that in turn to triggers required event. The adapted method for this approach allows user to flexibly and conveniently control multiple household appliances with simple gestures.

The outline of the paper is as follows. Section 2 presents an overview of smart home controller research. The proposed framework by using gesture based is described in Section 3. The performance evaluations are discussed in Section 4 and finally this paper is concluded in the last section.

2 Smart Home Controller

Efforts from researchers mainly focus on technologies and existing standards in integrating a centralized smart home controller that could control the connected devices and services. Results from this effort show that controller needs to be easy to handle, lightweight, and have more intuitive interface for people to interact with all the devices at home [4].

An investigation in [5] shows that people often prefer to use smart phones rather than the computers at home, as they are easily accessible and convenient to use. Moreover, touch screen based interaction is becoming prevalent in smart home. Utilizing these characteristics, various researches concentrates on creating smart home controller using smart phone or touch screen technologies. Some of the popular research can be seen in [4] called HouseGenie. The HouseGenie is a universal monitor and control of networked devices in smart home using common touch screen phone. By wirelessly communicating with an OSGibased portal, which maintains all the devices through varied protocols, HouseGenie facilitates universal home monitor and control. Figure 1 shows a glance of HouseGenie system.



Fig. 1. Glance of using HouseGenie in smart demo home based on (a) Android on Nexus One, and (b) Windows on Vivliv S5

Voice controller can also be applied in smart home environment. User could voice out command that allowed sound controller to distinguish the word, in turn triggering certain events. One such application can be seen in [6]; where the controller applied speech recognition library provided by SUNPLUS in identifying the command. The system captures the voice command pronounced by the user, parsing it and sends a control message to the home network that operate respective appliances.

A unique method has been proposed in [7] by using thoughts in controlling home appliance. By capturing certain level of electroencephalogram (EEG) or electrical brain activity, control commands could be given out in switching TV channels, opening and closing doors and windows, and navigation. EEG is captured by using special control masks that communicate with brain-computer interface (BCI) system. This method is currently tested only in smart home simulation but the possibility of applying it in real life may be realized in the future.

3 Gesture Based Controller (GaC) System Model

The gesture based controller (GaC) is a new method in controlling various household appliances by using simple and easily memorize gesture. GaC uses a glove based approach in recognizing the gesture produced by the user. In this system the user is required to wear a specially made glove that allows GaC to read commands from the user. The GaC framework can be divided into three main components as shown in Fig. 2; the gesture based recognition (GBR) system, data processing center (DPC) server and the household controller unit (HCU).

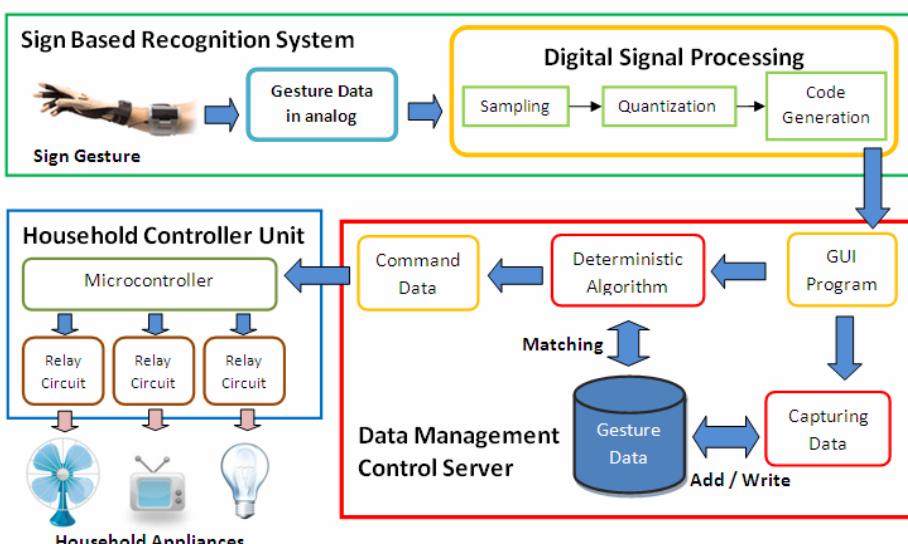


Fig. 2. The proposed framework of GaC for automating household appliances

The GBR system captures gestural input performed by the user and transforms it into digital data format or codeword. The digitized codeword data are then transmitted to the DPC server by using transceiver system. DPC performs a simple deterministic algorithm on the received codeword data to execute matching exercise. Once the matching exercise is determined, a control or command data is transmitted out to the HCU where required event is triggered.

3.1 Gesture Based Recognition (GBR) System

The GBR system is a wearable data glove system that recognizes gestures produced by a user or signer. It produces codeword data which represents the gesture produced by the signer. Gestural input are captured using flex sensors which are attached on top of each finger. The flex sensor is a resistance-varying strip that increases its total electrical resistance when the strip is flexed in one direction.

The function of the sensor is the same as a simple potentiometer capable of changing its resistance by twisting the potentiometer's knob. Voltage of the circuit is directly proportional to the resistance as given by Ohm as follows:

$$V \propto R \quad (1)$$

The varying changes of the resistor in turn change the voltage accordingly. By using the concept of voltage divider, V_0 represents the needed data for recognition purposes and that point is connected to the microcontroller as shown in Fig. 3. A total of three processes are done during digital signal processing (DSP); sampling, quantization and code generator. Fig. 4 displays the flow diagrams of DSP. In the process of sampling, the analog data are read at specific time interval. As for quantization, the obtained values are translated into discrete values. The maximum discrete value depends on the number of bits that can be supported by the microcontroller. The code generator process is a process that converts the quantized value into appropriate data value for transmission. Once the conversion process has been done, the codeword data is transmitted to the DPC through a transceiver system. The Bluetooth system has been selected as the transceiver system for this experiment.

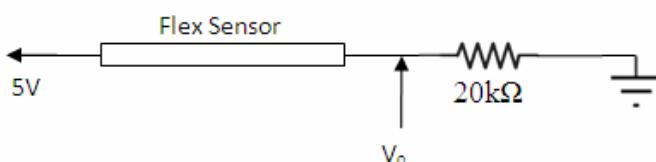


Fig. 3. Flex sensor interfacing with the microcontroller

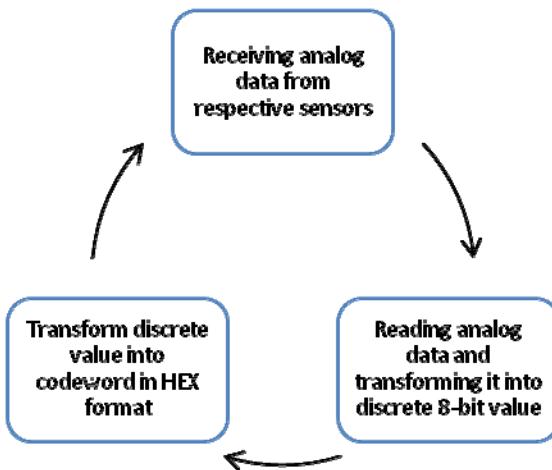


Fig. 4. The DSP flow diagram of GBR system implementation

3.2 Data Processing Center (DPC)

The data processing center (DPC) is resided on a standalone server or a computer to provide deterministic process in recognizing and determining the codeword data. A graphic user interface (GUI) program has been developed that allow user to select two of the main functions of the system; the activation function and the edit function. The activation function provides the necessary command data that is corresponding to the codeword data produced by the user. A simple matching algorithm is used on the codeword data and gestural data to produce the required command data. Matching is achieved when the codeword data is matched with one of the gestural data stored in the database. The produced command data contains the necessary information required in triggering events to the household appliances. The edit function allows user to redefine the gestural data with a new one. This is done by capturing the new gestural data and replacing it with the old gestural data.

3.3 Home Controller Unit (HCU)

The home controller unit (HCU) plays the role of controlling the household appliance according to the received command data from DPC server. By combining microcontrollers and relay circuits, simple on/off function can be performed on the electrical appliances. The relay circuit is turned ON when it receives a high input from the microcontroller and turned OFF upon receiving a low input. Moreover, communication between the HCU and DPC server is done using either serial or wireless communication. Again, Bluetooth is selected as the transceiver system in this set up.

4 Results and Discussion

The system is tested in a small scale environment consisting of five simple household appliances. Each household appliance requires two command data to operate a simple on/off scenario. Therefore, a total of ten sets of gestural data are needed in this experiment. Before the experiment begins, DPC is first ‘trained’ to recognize the ten sets of gestural data. Fig. 5 shows ten examples of command data that can be used by a single glove.



Fig. 5. Example of possible gestural command data

As seen in Fig. 5, each gestural data is unique to each other. The system is designed as such that each distinct gestural data is corresponding to a command data that could control the household appliances. In order to read the gestural input, users (or signers) must wear GBR as shown in Fig. 6.



Fig. 6. Gestural based recognition (GBR) system with developed (a) data glove and (b) control circuit

DPC allows user to ‘train’ the system by selecting the edit function as shown in Fig. 7. By selecting the corresponding command, a user could train the system using any gestural input. Once the ‘training’ is complete, the system is ready to be used by selecting the activation function from DPC.

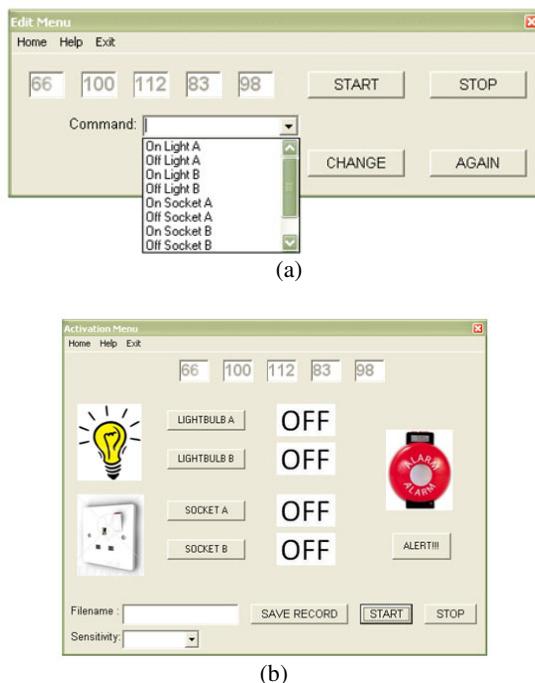


Fig. 7. A developed GUI in DPC server with (a) the edit function and (b) the execution / activation function

The command data is produced by DPC by using deterministic approach where a matching algorithm is done on the codeword data. An example can be seen in Fig. 8 where two examples of gestural data are used in controlling one of the household appliances.

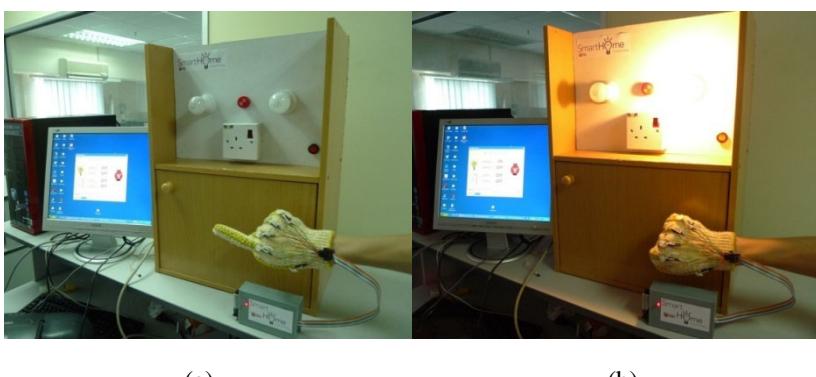


Fig. 8. Operation of the system where the light bulb is (a) turned 'OFF' and (b) turned 'ON'

In addition to the executing of the required function, the system has been tested on its flexibility in allowing user to change the gestural data. This is done by capturing the new gestural input produced by user and stores it in the database, replacing the old gestural data. This characteristics and ability allows user to freely and flexibly change the gestural data into one that is recognizable or easily produced by the user. It is helpful to user especially disabled people that could not perform certain gestures freely. Fig. 9 shows an example of changing a corresponding gestural input to a new input.

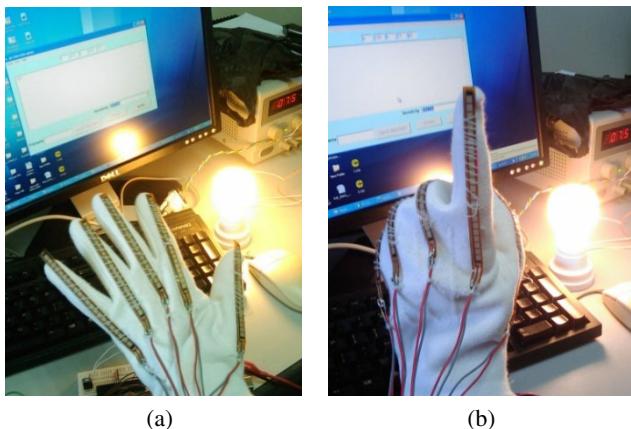


Fig. 9. Examples of editing gestural input of light bulb from sign (a) to sign (b)

The proposed system offers a new approach in smart home controller system by utilizing gesture as a remote controller. The adapted method for this innovative approach allows user to flexibly and conveniently control multiple household appliances with simple gestures. Additionally, the system is built using low-cost approach that could be implemented easily onto any household.

Overall, the whole system is not a perfect controller as it comes with a number of limitations. One of the hurdle is determining when the user is addressing the system; especially the GBR system. This skill is beyond the current sign based recognition glove systems that is available in the market. Therefore, the application program must allow users to inform the system that the GBR is being addressed. Additionally, the system must be smart enough to determine when the system is idle when not in used.

Additionally, the GBR and DPC system requires additional improvement on the recognition algorithm. Currently, the system implements a simple matching algorithm which could only support simple gestural recognition. For advance recognition, it is required to have better sensory and algorithm.

Furthermore, the system can be further improved by eliminating the need of data glove. This can be substituted by using visual-based gesture translation system that provides gloveless system. As its name indicated, visual-based system that uses visual recognition can be employed in acquiring the expression and gestural movement of the signer. It requires three main stages in translating the gesture; video acquisition, video processing and lastly the translation process. Some examples of visual-based translation can be seen in [8], [9].

5 Conclusion

This paper presents a new and alternative approach in providing controller to a smart home system using gesture. The proposed method uses a data glove to translate and capture gestural input of the user. By using deterministic approach in translating and parsing the gestural input, user could control from simple to complex home appliances. Additionally, the system has been designed to be flexible, in which it allows user to freely change the gestural input corresponding to a command data. However, the proposed scheme comes with certain limitations i.e. to determine when the system is being addressed and sensitivity issue. Improvements can be further done by introducing visualization instead of using data glove. Future works would include improving its accuracy and sourcing cheaper methods in capturing the gestural data.

References

1. Chan, M., Hariton, C., Ringeard, P., Campo, E.: Smart House Automation System for the Elderly and the Disabled. In: IEEE International Conference on Systems, Man and Cybernetics, Intelligent Systems for the 21st Century, vol. 2, pp. 1586–1589 (1995)
2. Dario, P., Micera, S., Macri, G., Carpaneto, J., Carrozza, M.C.: Biorobotics for Longevity. In: 9th International Conference on Rehabilitation Robotics (ICORR 2005), pp. 269–272 (2005)
3. Hwang, I.K., Lee, D.S., Baek, J.W.: Home Network Configuring Scheme for All Electric Appliances using Zigbee-Based Integrated Remote Controller. *IEEE Transactions on Consumer Electronics* 55(3), 1300–1307 (2009)
4. Suo, Y., Wu, C., Qin, Y., Yu, C., Zhong, Y., Shi, Y.: HouseGenie: Universal Monitor and Controller of Networked Devices on Touchscreen Phone in Smart Home. In: 7th International Conference on Ubiquitous Intelligence & Computing and 7th International Conference on Autonomic & Trusted Computing (UIC/ATC), pp. 487–489 (2010)
5. Nylander, S., Lundquist, T., Brännström, A., Karlson, B.: “It’s Just Easier with the Phone” – A Diary Study of Internet Access from Cell Phones. In: Tokuda, H., Beigl, M., Friday, A., Brush, A.J.B., Tobe, Y. (eds.) *Pervasive 2009. LNCS*, vol. 5538, pp. 354–371. Springer, Heidelberg (2009)
6. Guo, J.K., Lu, C.L., Chang, J.Y., Li, Y.J., Huang, Y.C., Lu, F.J., Hsu, C.W.: Interactive Voice-Controller Applied to Home Automation. In: Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, pp. 828–831 (2009)
7. Holzner, C., Guger, C., Edlinger, G., Gronegress, C., Slater, M.: Virtual Smart Home Controlled by Thoughts. In: 18th IEEE International Workshops on Enabling Technologies: Infrastructures for Collaborative Enterprises, pp. 236–239 (2009)
8. Akmeliawati, R., Ooi, M.P.L., Kuang, Y.C.: Real-Time Malaysian Sign Language Translation using Colour Segmentation and Neural Network. In: IEEE Instrumentation and Measurement Technology Conference Proceedings, pp. 1–6 (2007)
9. Yang, Q., Peng, J.: Chinese Sign Language Recognition for a Vision-Based Multi-features Classifier. In: International Symposium on Computer Science and Computational Technology, vol. 2, pp. 194–197 (2008)

Upper Body Gesture Recognition for Human-Robot Interaction

Chi-Min Oh, Md. Zahidul Islam, Jun-Sung Lee, Chil-Woo Lee, and In-So Kweon

Chonnam National University, Korea,
Korea Advanced Institute of Science and Technology, Korea
{sapeyes, zahid, aliesim}@image.chonnam.ac.kr,
leecw@chonnam.ac.kr,
iskweon@kaist.ac.kr

Abstract. This paper proposes a vision-based human-robot interaction system for mobile robot platform. A mobile robot first finds an interested person who wants to interact with it. Once it finds a subject, the robot stops in the front of him or her and finally interprets her or his upper body gestures. We represent each gesture as a sequence of body poses and the robot recognizes four upper body gestures: “Idle”, “I love you”, “Hello left”, and “Hello right”. A key pose-based particle filter determines the pose sequence and key poses are sparsely collected from the pose space. Pictorial Structure-based upper body model represents key poses and these key poses are used to build an efficient proposal distribution for the particle filtering. Thus, the particles are drawn from key pose-based proposal distribution for the effective prediction of upper body pose. The Viterbi algorithm estimates the gesture probabilities with a hidden Markov model. The experimental results show the robustness of our upper body tracking and gesture recognition system.

1 Introduction

For a long time, gesture recognition has been a practical and useful research for human-robot interaction (HRI). Implementing the gesture recognition method, a robot system firstly has to segment the pose of user action and then infer a meaning of the sequential pose as a gesture.

The pose recognition can perform in two ways. The first way is so called feature-based method [1] and it obtains the visual features with about pose manifolds from an image sequence. Many interesting algorithms such as principle component analysis (PCA) and independent component analysis (ICA) have been applied for feature analysis, and nonlinear and strong pattern classifiers such as multi-layer perceptron (MLP) and support vector machine (SVM) have been used for pose recognition. However, feature-based method has limitations of generosity; it cannot be applied for arbitrary poses which did not participate in training procedure, and the poses of different people not adjusted to the body structure of a specific-sized person.

By contrast, the second way is model-based method [2,3,4] and it adjusts to any pose of inter persons with a 2D parts-based body model which is made for a specific-sized person but can accept the some variance of inter-person poses. This model

describes the body pose as a configuration vector. Estimating a correct configuration vector of the body model in an image determines the pose of an acting user but is nontrivial due to the high degree of freedom (DOF). Therefore, some of robust tracking methods are used; so called linear tracking method, the Kalman filter, as in [5,6], but results in inconsistent tracking for pose estimation. Nonlinear tracking method, the particle filter, as in [7,8] have been applied for this issue. Adopting particle filter can cover the problem of high DOF but the computational cost increases because it needs the great number of particles (hypotheses) to cover the high DOF.

Most of tracking methods usually predict multiple hypotheses based on the assumption of Markov process which is the critical reason for creating many particles. To overcome the problem of high DOF with Markov chain model, key poses can be one of the efficient clues for prediction with smaller particles and we assume that the similarities between key poses and the input image can be a better proposal distribution, as in [9] where we usually draw particles in the prediction step of particle filtering.

This paper proposes a new proposal distribution which is based on key poses and Makov chain model which is still used to recycle the good particles from the previous tracking step. By using both ways for the proposal distribution, we track upper body poses robustly and obtain the pose sequence as a result. In the pose sequence, each tracked pose is named with pose number determined from the number of the closest key pose. For gesture recognition we use HMM which represents the relationship between gesture states and pose events. We use four gestures already explained before, so this means four gesture states are defined in HMM, however, we cannot understand the gestures directly from body tracking. We need to transform the pose sequence into a gesture sequence using HMM and the Viterbi algorithm.

This paper is organized as follows: Section 2 describes our mobile robot platform for HRI. Thus, it describes the HRI operation with two steps: user selection and gesture recognition stages. In the user selection stage, from the omnidirectional view of its omni-camera, the robot is eager to find a person who wants to interact with the robot. In the gesture recognition stage, robot interprets the upper body motion as a gesture. After that, we explain the gesture recognition algorithm. Section 3 describes the model of upper body pose. Section 4 presents the upper body tracking algorithm using particle filtering. We define our key pose-based proposal distribution method in this section. Section 5 briefly describes the gesture recognition process. Section 6 shows the experimental results and section 7 concludes our work.

2 HRI in Mobile Robot Platform

In this work we use *NRLAB02 mobile platform* [10] which is developed by RedOne Initiative Inc. The aim of our work is to construct a prototype of an intelligent HRI system with visual interaction. Therefore the upper row of mobile robot is equipped with visual sensors and a computer. We installed two visual sensors: Microsoft's Kinect camera [11](TOF camera) and an ordinary Omni-camera as shown in Fig. 1. The Kinect camera is used to get the silhouette image of nearby people.

Our upper body tracking algorithm utilizes the silhouette image. The other camera, omni-camera implemented with a hyperbolic mirror made by Eizoh Inc., captures the panoramic images with the FOV of 360 degrees. The robot selects the interested person from the panoramic images.

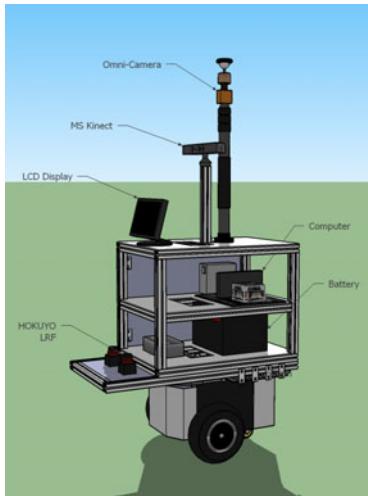


Fig. 1. Our mobile robot platform has omnidirectional camera for interaction-user selection and MS Kinect for gesture recognition

The scenario of HRI is as follows. At first, the robot finds the interested person who wants to interact with it. This is called “User Selection Stage”. Then, the robot interprets the intention of the person’s motion with the meaning of tracked body poses and we call this procedure “Gesture Recognition Stage”.

2.1 User Selection Stage

We assume that users are pedestrians appearing in panoramic images of omni-camera. HOG-SVM detector, as in [12] detects the nearby pedestrians but it is quite slow to find the pedestrians from a whole image. As pedestrians usually move around the robot, we assumed that moving area can possibly be the candidate areas of pedestrians to reduce the search area. However, the egomotion of the robot distorts the segmentation of moving region. So we use KLT tracker to estimate the egomotion of the robot [13] and obtain the egomotion-compensated frame difference. Therefore we can detect the pedestrians from only moving areas in a short time.

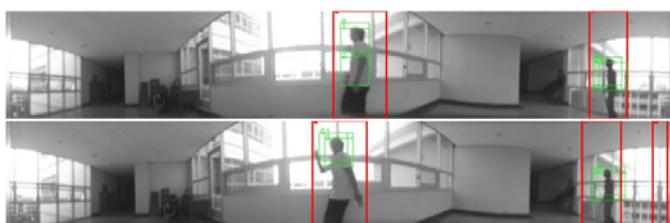


Fig. 2. Pedestrian detection in moving object areas

2.2 Gesture Recognition Stage

After user selection among pedestrians, the user can show the intention by upper body gestures. We recognize four upper body gestures: “Idle”, “I love you”, “Hello left” and, “Hello right”. The gestures are recognized from the pose sequence. To get the pose sequence we use our model-based upper body tracking. In this model-based approach, we define a 2D-parts-based upper body model.

3 Upper Body Model

We define an upper body model based on Pictorial Structures (PS) [2]. PS is known for an efficient method for tracking deformable objects. PS describes the deformable objects with 2D parts and each part is connected to its parent part with joint location, as shown in Fig. 3(b). Additionally the connected parts can be represented as a graphical structure, as shown in Fig. 3(c). Around joint locations all parts can translate and rotate. The benefit of PS model is that any pose of upper body can be made with 2D parts but 3D information is not regarded in PS. PS model can represent a sufficient number of 2D poses for gesture recognition. In addition to the benefit of PS, each pose of PS can be considered as a 2D template image. As shown in Fig. 3(a), by overlaying PS model on the silhouette image, we can simply measure the similarity between the silhouette image and 2D template image of PS model using frame difference or chamfer matching, as in [14].

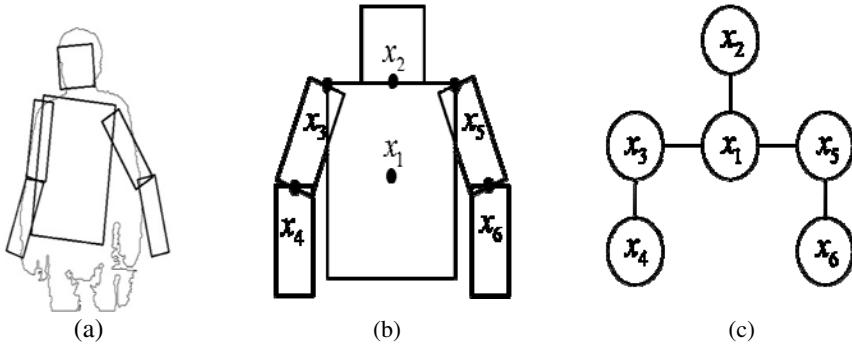


Fig. 3. The upper body model: (a) An overlaid model on a silhouette, (b) the joint locations of all parts, and (c) the representation of graphical relation between parts

Fig. 4 shows the body part of upper body model as $x_i = \{(x,y), (dx,dy), \theta, (w,h)\}$ parameters: joint location (x,y) , spring-like displacement (dx,dy) , orientation θ and rectangular sizes (w,h) . Among the parameters, the joint location (x,y) cannot change by its part and only can change by parent part’s location. This constraint makes each part connected with its parent. On the other hand, the only way of moving around its parent is related to (dx,dy) displacement vector. This is so called spring-like displacement determined by Gaussian distribution with the mean (x,y) and a variance which is a control parameter. In addition to the displacement, the part can rotate with the orientation parameter θ based on the rotation origin, (x,y) .

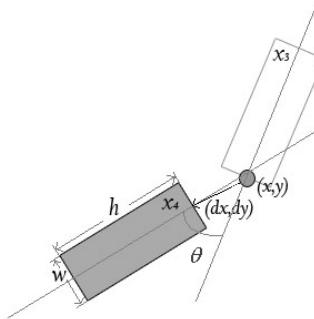


Fig. 4. The parameters of child part x_4 which is connected to parent part x_3

4 Upper Body Tracking

4.1 Particle Filtering

The assumption of most tracking methods is that the current information must be similar to and dependent on the previous information at each time step. Particle filter estimates a proper posterior distribution by updating the posterior distribution at previous tracking time step. The proposal distribution predicts the current posterior distribution from previous posterior distribution with discrete and weighted particles. In Eq. (1), the posterior distribution $p(x_{t-1} | y_{t-1})$ at previous time step represents itself with discrete and weighted particles, where $x_{t-1}^{(i)}$ is i th particle, $w_{t-1}^{(i)}$ is the weight of i th particle, and N is the total number of particles.

$$p(x_{t-1} | y_{t-1}) \approx \{(x_{t-1}^{(i)}, w_{t-1}^{(i)})\}_{i=1}^N \quad (1)$$

As in [8], particle filter has two steps: prediction and update. In the prediction step, the previous posterior distribution is marginalized to eliminate x_{t-1} and to be updated to x_t based on transition model $p(x_t | x_{t-1})$, Markov chain model.

$$p(x_t | y_{t-1}) = \int p(x_t | x_{t-1}) p(x_{t-1} | y_{t-1}) dx_{t-1} \quad (2)$$

In the update step, the posterior distribution is reformulated to adjust to the current observation y_t . Based on Baye's rule, posterior distribution $p(x_t | y_t)$ is represented with the likelihood $p(y_t | x_t)$ and prior distribution $p(x_t | y_{t-1})$.

$$p(x_t | y_{t-1}) = \frac{p(y_t | x_t) p(x_t | y_{t-1})}{p(y_t | y_{t-1})} \quad (3)$$

In addition to the prediction step, the particles for x_t prediction are drawn from the proposal distribution $q(x_t | x_{t-1}, y_t)$ and in the update step the weights of particles are determined by below:

$$w_t^{(i)} = w_{t-1}^{(i)} \frac{p(y_t | x_t^{(i)}) p(x_t^{(i)} | y_t)}{q(x_t^{(i)} | x_{t-1}^{(i)}, y_t)} \quad (4)$$

In the process of weighting particles, likelihoods of particles are measured. Our likelihood $p(y_t | x_t^{(i)})$ is a joint likelihood with edge and silhouette likelihoods.

$$p(y_t | x_t^{(i)}) = p(I_S | x_t^{(i)}) p(I_E | x_t^{(i)}) \quad (5)$$

$p(I_S | x_t^{(i)}) = \exp(-\|I_S - I_{SM}\|)$ is the likelihood of silhouette matching and $p(I_E | x_t^{(i)}) = \exp(-d(I_E, x_t^{(i)}))$ is the likelihood of chamfer matching as in Fig. 5.

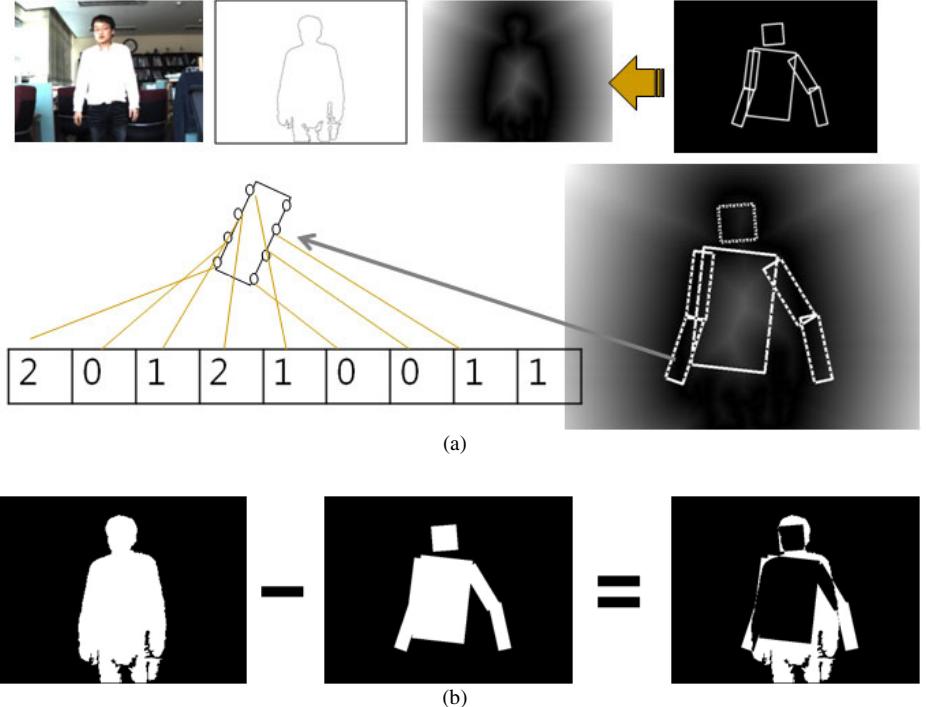


Fig. 5. The joint likelihood of (a) silhouette likelihood and (b) edge likelihood

4.2 Key Pose-Based Proposal Distribution

As mentioned in introduction section, the formal proposal distribution with first order Markov chain, $q(x_t^{(i)} | x_{t-1}) = p(x_t^{(i)} | x_{t-1})$, could not cover abrupt motion. The occurrence of abrupt motion is mainly up to the time leaping of the slow system. The increased particles for covering high DOF are the reason of low speed. Therefore, key poses are useful for prediction of abrupt motion and reducing the number of particles.

We define a key pose library (KPL) which consists of 13 key poses. Fig. 6 shows all the key poses as PS models and silhouette images. Each key pose is represented as $k_i = (I_i, PS_i, f_i)$; I_i is the pose image of key pose, PS_i is PS model, and f_i is visual feature of key pose, as in [9]. Visual features are used for measuring key pose similarity. Based on the KPL, the key pose-based proposal distribution is defined as

$$q(x_t | x_{t-1}, y_t, KPL) = \alpha p(x_t | x_{t-1}) + (1-\alpha) p(x_t | y_t, KPL). \quad (6)$$

This proposal distribution is a combined version of Markov chain-based proposal distribution and KPL-based prediction model which is defined as

$$p(x_t | y_t, KPL) = \frac{1}{13} \sum_{k=1}^{13} p(x_t | PS_k) p(y_t | PS_k). \quad (7)$$

KPL-based prediction model predicts the particles to be similar to some of key poses in KPL. The latest observation y_t is referred in KPL-based prediction model to measure how much key poses are similar to the current observation using $p(y_t | PS_k)$. $p(x_t | PS_k)$ is the probability for similarity to key pose k of the predicted state x_t .

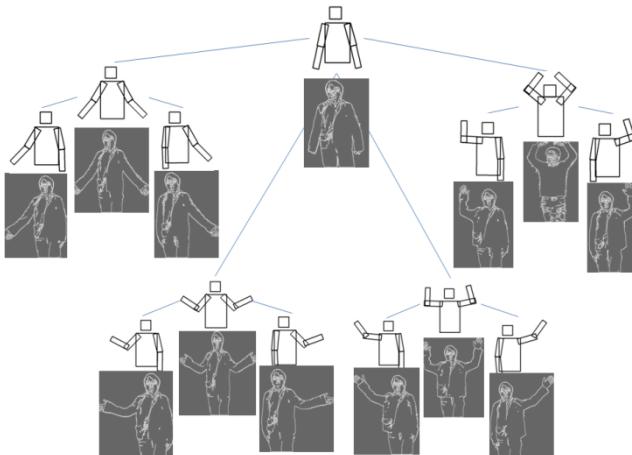


Fig. 6. Key Pose Library with 13 key poses (PS and Silhouette)

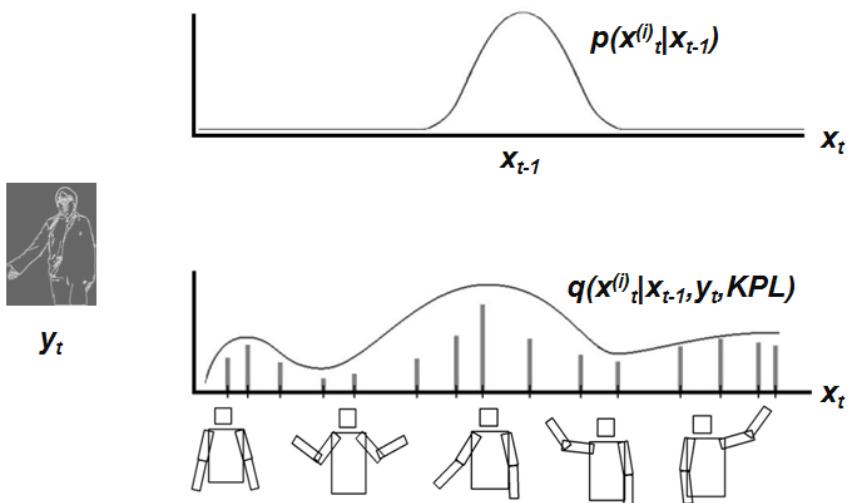


Fig. 7. Conceptual difference between our proposal distribution and Markov chain model

In Fig.7, we conceptually describe the difference of two proposal distributions. Consequently, our proposal distribution benefits from sparsely selected key poses which make it possible to predict particles globally. On the contrary, Markov chain model-based proposal distribution locally predicts particles around the previous state. When abrupt motion happens, our proposal distribution will predict better.

5 Gesture Recognition

From the pose sequence, we recognize the upper body gestures. We have 4 gestures: “Idle”, “I Love You”, “Hello Left” and “Hello Right”. The gesture recognition considers poses as events in the HMM structure. The events are seeable results of the pose tracking system but gesture states cannot be directly estimated from the system. HMM defines the gesture state transition matrix and the event emission matrix. Using Viterbi algorithm and HMM, we obtain the probability of hidden gesture states.

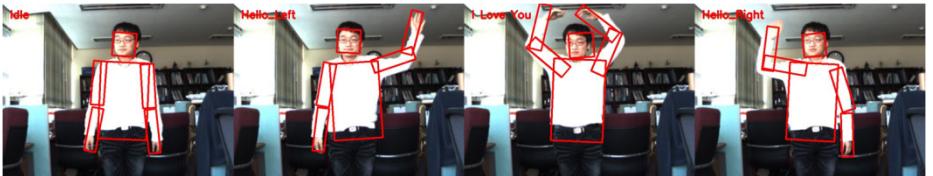


Fig. 8. Upper body gestures: “Idle”, “Hello Left”, “I Love You”, and “Hello right”

6 Experimental Result

For evaluation, we use our ground-truth database. By comparing with bootstrap, we found that our key pose-based proposal distribution overcomes the weakness of Markov chain model. Markov chain model has affected the results getting slow and failed from abrupt motion. With only 100 particles, KPL efficiently predicts smaller particles and overcomes the failures from abrupt motion, as shown in Fig. 9. Fig 10 shows the evaluated resulting images with several people. Our system adaptively tracks the different genders and heights.

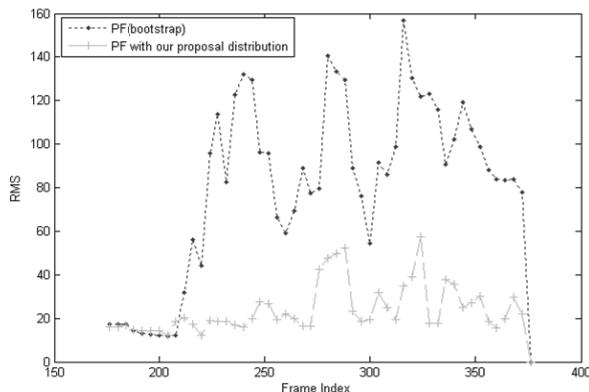


Fig. 9. Root Mean Square Error of bootstrap and our method

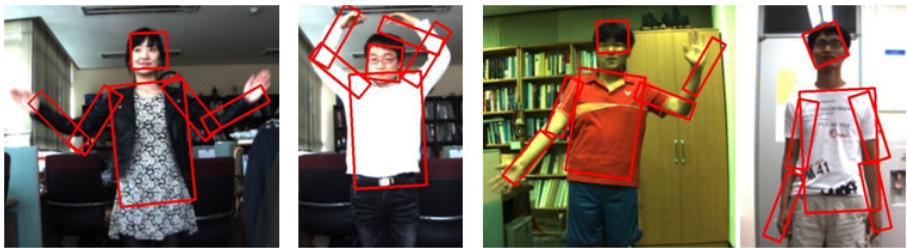


Fig. 10. Tracking results from several people

7 Conclusion

We have argued that key poses can be useful to improve the upper body tracking. As shown in Fig. 9, our system outperformed with only 100 particles using key poses. In prediction of particles and in recovering errors, key pose-based proposal distribution has been a great role in our system.

In the future, we will investigate how the key poses are mathematically working in particle filtering framework. Additionally we will try to build a regressive model of KPL to cover whole pose spaces not only 13 key poses. Finally we will test our system in the developing mobile robot platform.

Acknowledgement

This research was supported by the MKE(The Ministry of Knowledge Economy), Korea, under the Human Resources Development Program for Convergence Robot Specialists support program supervised by the NIPA(National IT Industry Promotion Agency) (NIPA-2010-C7000-1001-0007).

References

- [1] Thurau, C., Halavac, V.: Pose Primitive based Human Action Recognition in Videos or Still Images. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (2008)
- [2] Felzenszwalb, P., Huttenlocher, D.: Pictorial Structures for Object Recognition. International Journal of Computer Vision 61(1), 55–79 (2005)
- [3] Oh, C.M., Islam, M.Z., Lee, C.W.: A Gesture Recognition Interface with Upper Body Model-based Pose Tracking. In: International Conference on Computer Engineering and Technology, vol. 7, pp. 531–534 (2010)
- [4] Andriluka, M., Roth, S., Schilele, B.: Pictorial Structures Revisited: People Detection and Articulated Pose Estimation. In: International Conference on Computer Vision and Pattern Recognition, pp. 1014–1021 (2009)
- [5] Barker, A.L., Brown, D.E., Martin, W.N.: Bayesian Estimation and the Kalman Filter. Computer & Mathematics with Applications 30(10), 55–77 (1995)
- [6] Weng, S.K., Kuo, C.M., Tu, S.K.: Video Object Tracking using Adaptive Kalman Filter. Journal of Visual Communication and Image Representation 17(6), 1190–1208 (2006)

- [7] Merwe, R.A., Freitas, N.D., Wan, E.: The Unscented Particle Filter. In: Advances in Neural Information Processing Systems, vol. 13 (2001)
- [8] Islam, M.Z., Oh, C.M., Lee, C.W.: Real Time Moving Object Tracking by Particle Filter. In: International Symposium on Computer Science and Its Application, pp. 347–352 (2008)
- [9] Oh, C.M., Islam, M.Z., Lee, C.W.: Pictorial Structures-based Upper Body Tracking and Gesture Recognition. In: Korea-Japan Joint Workshop on Frontiers of Computer Vision (2011)
- [10] RedOne Technology (February 2011), <http://urc.kr/>
- [11] MS Kinect (February 2011), <http://www.xbox.com/en-US/kinect>
- [12] Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian Detection: A Benchmark. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (2009)
- [13] Oh, C.M., Setiawan, N.A., Aurahman, D., Lee, C.W., Yoon, S.: Detection of Moving Objects by Optical Flow Matching in Mobile Robots using an Omnidirectional Camera. In: The 4th International Conference on ubiquitous Robots and Ambient Intelligence (2007)
- [14] Barrow, H.G., Tenenbaum, J.M., Bolles, R.C., Wolf, H.C.: Parametric Correspondence and Chamfer Matching: Two New Technique for Image Matching. In: The 5th International Joint Conference on Artificial Intelligence, pp. 1175–1177 (1997)

Gaze-Directed Hands-Free Interface for Mobile Interaction

Gie-seo Park, Jong-gil Ahn, and Gerard J. Kim

Digital Experience Laboratory

Korea University, Seoul, Korea

{2002, hide989, gjkim}@korea.ac.kr

Abstract. While mobile devices have allowed people to carry out various computing and communication tasks everywhere, it has generally lacked the support for task execution while the user is in motion. This is because the interaction schemes of most mobile applications are centered around the device visual display and when in motion (with the important body parts, such as the head and hands, moving), it is difficult for the user to recognize the visual output on the small hand-carried device display and respond to make the timely and proper input. In this paper, we propose an interface which allows the user to interact with the mobile devices during motion without having to look at it or use one's hands. More specifically, the user interacts, by gaze and head motion gestures, with an invisible virtual interface panel with the help of a head-worn gyro sensor and aural feedback. Since the menu is one of the most prevailing methods of interaction, we investigate and focus on the various forms of menu presentation such as the layout and the number of comfortably selectable menu items. With head motion, it turns out 4x2 or 3x3 grid menu is more effective. The results of this study can be further extended for developing a more sophisticated non-visual oriented mobile interface.

Keywords: Mobile interface, Gaze, Head-controlled, Hands-free, Non-visual interface.

1 Introduction

Mobile interaction has become an important issue with the explosive usage of hand-held devices. Most mobile interfaces are suited only for the situation where the user is stationary and interacting closely (visually) and stably with the device in hand. However, there are also many occasions where one needs to interact with the mobile device while in motion. In this situation, it is often difficult to make out (let alone use) the visual display and directly interact with the objects (on the touch screen or hardware buttons) effectively.

On the other hand, sensors external to the hand-held mobile device, but integrated with embedded in the body accessories (e.g. eye glasses, headset, ear piece, hat, necklace) are becoming popular and opening new opportunities in mobile interaction area.

In this paper, we propose an interface which allows the user to interact with the mobile devices during motion without having to look at it (non-visual) or use one's hands (hands-free). Being non-visual and hands-free, the user is allowed to carry on the main motion task without visual distraction. Specifically, our interface proposal is to use gaze/head motion gestures, with an invisible virtual interface panel with the help of a head-worn gyro sensor and aural feedback. Furthermore, as the menu is one of the most prevailing methods of interaction, we investigate and focus on the various forms of menu presentation, for the proposed interface, such as the most effective and usable layout and the number of comfortably selectable menu items.

This paper is organized as follows: a brief description of related research is given in the next section. Section 3 provides the implementation details of the proposed interface. The menu design space exploration experiment is presented in Section 4 with the results discussed in Section 5. Finally we conclude the paper with an executive summarization of our contribution and directions for future work.

2 Related Work

Recent smart phones are equipped with many sensors such as the GPS, tilt sensor, accelerometer, digital compass, camera, etc. In relation to our work, many interfaces taking advantage of these sensors have appeared, e.g. for pointing and scrolling [14], menu marking [13] changing screen orientation [8], zooming and navigation [5] and shaking based event enactment [17]. However, most of these attempts were still centered on the mobile device itself and require visual recognition of the interaction mode and results.

For eyes-free operation, Oakley and O'Modhrain [12] proposed a tile based interaction with tactile feedback for menu navigation. Brewster et al. also developed an eyes-free interface that used 3D aural cues for choosing menu items [2]. While eyes-free, note that these interfaces still require the interaction to be applied in a stable non-moving condition.

With hands occupied, the human head can act as a helpful input medium, providing directional or gestural cues. Head/gaze based interaction has been applied with mixed results for non-mobile computing environments [1][18], but to a less extent for mobile situations. Crossan et al. [3] investigated the possibility to use head tilting for mobile interaction during motion. While their interface still required the use of hands and eyes, our work was inspired by their approach.

3 Interface Design

3.1 Basic Idea: Gaze and Sound Directed

The basic idea is to build a head-controlled and hands-free interface so as to select a menu item while a user is in motion and hands occupied (walking, running or driving). As an application is started, the existence of a virtual pop-up menu is notified to the user via voice and enters the menu selection mode. The user is instructed to first center one's head. Then, a virtual ray emanating from the user's head, in the direction of one's gaze, to a virtual menu panel/item situated in the front.

Since the menu is virtual, interaction with the menu system is guided through sound and voice (e.g. which menu items are where). The gaze is used to select a menu item and the final confirmation is given by shaking/nodding one's head or by gazing for a fixed amount of time, also accompanied with a final aural feedback (Figure 1).

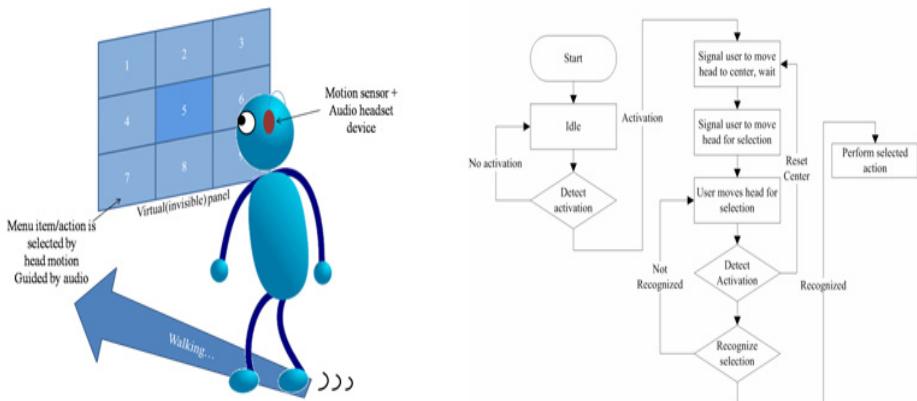


Fig. 1. The proposed gaze and sound directed menu interface

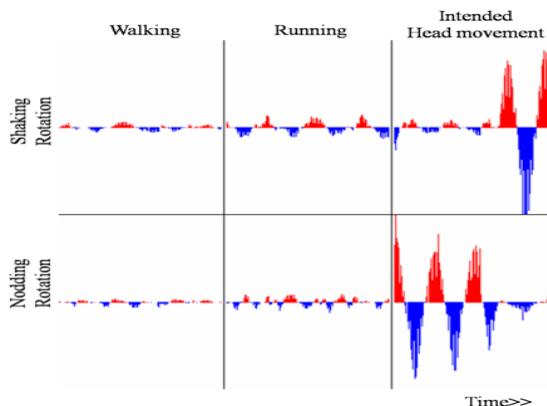


Fig. 2. Sensor values of the head-attached gyrometer for intentional nodding or shaking over walking or running motions

The head movement and direction is detected by a wirelessly operated (Bluetooth) gyro sensor worn on the head (note that even if the sensor is not attached to the main mobile device, such a configuration is deemed quite acceptable and not so inconvenient from the example of already popular Bluetooth operated headsets). A gyro sensor was used (e.g. over an accelerometer) due to the easier discerning between true head motion and mere body motion. Figure 2 shows that indeed the

intended nod or shaking was easily detected over the running or walking motion. In actual implementation for the experiment, we attached the G2 Wireless Mouse (from Ocean Technology, which uses the MG1101 dual axis gyrometer from the Gyration Corporation) to a cap (see Figure 3).



Fig. 3. Wireless gyrometer sensor attached to a cap

The virtually selectable menu panels can vary in several aspects, e.g. in its form, the use of gestural selection method (as an alternative to gaze based selection, Figure 4), confirmation method (Figure 5), etc. Note that the selection and confirmation methods are coupled with the form of the menu. For instance, with 2D menus, selection can be made by moving over the panel in “4-ways” and making a “Round trip” gaze movement for confirmation. Such a method would not be possible with a 1D menu.

The numbers of menu items is set by equally dividing user’s effective head angle range. Human head normally can be rotated about 140 degrees for shaking and 100 degrees for nodding [9]. However, rotating the head in such a wide range is practically not feasible, since the user needs to attend to the other on-going task.

For this reason, calibration process for a user to measure the neck rotating range was performed. The range was detected while a user rotates head with keeping the vision for moving activity.

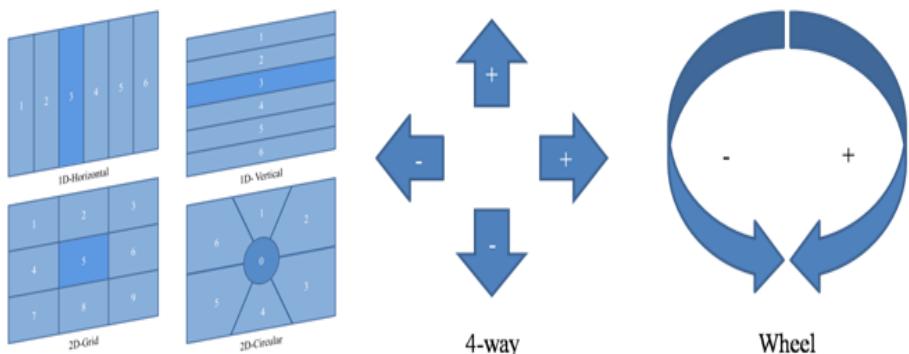


Fig. 4. Various forms of virtual menu and examples of gestural methods for menu selection (as an alternative to gaze based selection)

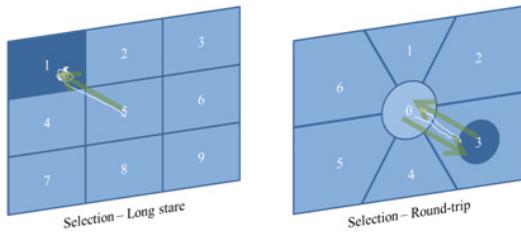


Fig. 5. Two examples of confirmation method: “Long stare” vs. “Round trip”

4 Experiment

A usability experiment was carried out to find the preferred form of the virtual menu and the proper number of effectively selectable menu items (in a fixed area).

4.1 Experiment Design

The two factors in the experiment was the type or form of the menu and the number of menu items. This resulted in about total of 19 treatment groups. For all treatments, gaze based selection was used and both “Long-stare” and “Round-trip” based confirmation methods were tested (see Figure 5).

Table 1. 19 tested combinations between the type of menu and no. of menu items

Menu Form	No. of Menu Items
1D Horizontal	2, 3, 4, 5, 6, 7, 8
1D Vertical	2, 3, 4, 5, 6, 7, 8
2D Grid	4 (2x2), 6 (3x2), 8 (4x2), 9 (3x3), 4x3 (12)

The user was given a menu selection task and the total accumulated selection time was used as the main dependent variable reflecting of the user performance. Selection error was also recorded. To help precise selection, the selected item was enlarged (see Figure 6). Even though the enlarged (twice the size of the menu item) virtual menu item is not visible, the enlarged size helps the user stay with the menu item within a tolerable range and makes the interaction robust to noise and nominal user shaking.

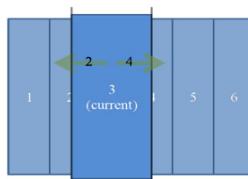


Fig. 6. Increasing the tolerance for a selected menu item (for reducing error and making the system robust to sensor/motion noise)

4.2 Experiment Process

The experiment proceeded in two stages: first with the 1D menus to find the upper limit for the number of menu items for 2D menus in each horizontal and vertical dimension. 1D horizontal and vertical menus with 2 to 8 items (smaller relative menu item width with increasing number of items) were tested first. The experiment was conducted in motion on a tread mill with the speed set at 4km/hour.

Based on this experiment, the upper limit for the 2D menu items were set at 4x3v. 2D menus with 4x3, 4x2, 3x3, 3x2 and 2x2 configurations were tested in the second stage of the experiment. Each experiment was run as a two factor within-subject repeated measure. During the experiments, the subject was given a particular menu configuration in a balanced order and asked to carry out a menu selection task, three times. The first trials and extreme outliers were excluded from the analysis. The subject's task performance was recorded in terms of the accuracy and time of completion. A general usability survey was also taken afterwards.

5 Experiment Results

5.1 1D Menu

The experiment results (proper number of menu items) for the 1D (horizontal and vertical) menus are shown in Table 2 and Figure 7. The results show that as the menu item increases (and thus the relative size of the menu items decrease), the selection time increases and the accuracy is reduced (obviously). However, in general, the horizontal 1D menu is faster and more accurate, despite the slightly longer operation range (than that of the vertical 1D menu given the same number of menu items/divisions). The 1D horizontal menu is also more stable (less deviation).

Table 2. The selection time and accuracy for 1D horizontal and vertical menus

Division		8	7	6	5	4	3	2
(L/R shake)	Accuracy (%)	66.7	80.0	100.0	100.0	100.0	100.0	100.0
	Average (sec)	5.8	5.2	4.6	3.6	3.2	3.1	2.6
	StD (sec)	0.6	0.2	0.2	0.2	0.1	0.2	0.03
(U/D nod)	Accuracy (%)	0.0	66.7	72.7	88.9	100.0	100.0	100.0
	Average (sec)		6.1	5.5	4.8	4.0	2.7	2.6
	StD (sec)		1.5	1.1	0.8	0.6	0.06	0.06

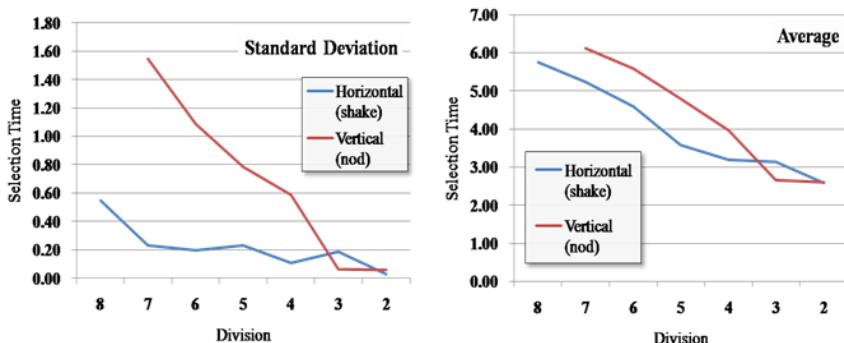
5.2 2D Menu

Based on the 1D result, the maximum grid division for the 2D menus was decided to be 4x3 (e.g. average selection time < 4 seconds and standard deviation < 0.2 seconds). Starting from 4x3, we reduced the 2D menu configuration down to 2x2 in the experiment.

For 2D menus, 4x2 and 3x3 grid menus proved to be the optimal configurations (in consideration of the selection time and associated accuracy). Interestingly, the number of menu items (8-9) coincided with the capacity of the human short term memory [11]. Table 3 shows the experimental results (selection accuracy and average task completion time/std.).

Table 3. The selection time and accuracy for 2D menus

Division	4x3	4x2	3x3	3x2	2x2
Accuracy (%)	88.9	100.0	100.0	100.0	100.0
Average (sec)	4.9	3.4	3.4	3.5	3.2
StD (sec)	0.7	0.2	0.2	0.2	0.3

**Fig. 7.** 1D menu results for selection time (standard deviation and average)

5.3 Circular Menus

In a separate/follow-up experiment, 2D circular menus were also tested. For the same number of menu items, compared to grid type, the circular menus did not perform well especially accuracy-wise. This was due to the slight confusion with the number layout being different from the clock. Also the “Round trip” style of menu item confirmation resulted in slower performance and high error (in having to deal with two spatial locations).

5.4 ANOVA Result

Analysis of variance (ANOVA) was applied to the experimental data (total selection time) and the main effect of the different menu forms/sizes was verified (p -value < 0.000 and F -value = 24.266) as well. The post-hoc test also showed such a similar trend and grouping among the menu types/sizes (we omit the detailed statistic values).

Table 4. Analysis of variance (total selection time)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	244.29	21	11.63	24.27	.000
Within Groups	80.54	168	.479		
Total	324.83	189			

5.5 Fitt's Law

The resulting performance may be explained by the Fitt's law [6] even though the proposed interface uses gaze and head movement, instead of the hand as was [15], to move the cursor/object and accomplish the given task. According to the Fitts' Law and other variations [10][16] the movement time (task time, MT) to select a target of width W at the distance D is related as follows:

$$MT = a + b \log_2\left(\frac{D}{W} + c\right)$$

where a, b and c are constants induced by a linear regression. The value inside the log function is also called the index of difficulty (ID), and ID is to exhibit a linear relationship with the MT. The experimental data, as shown in Figure 8, generally adhered to the Fitt's Law.

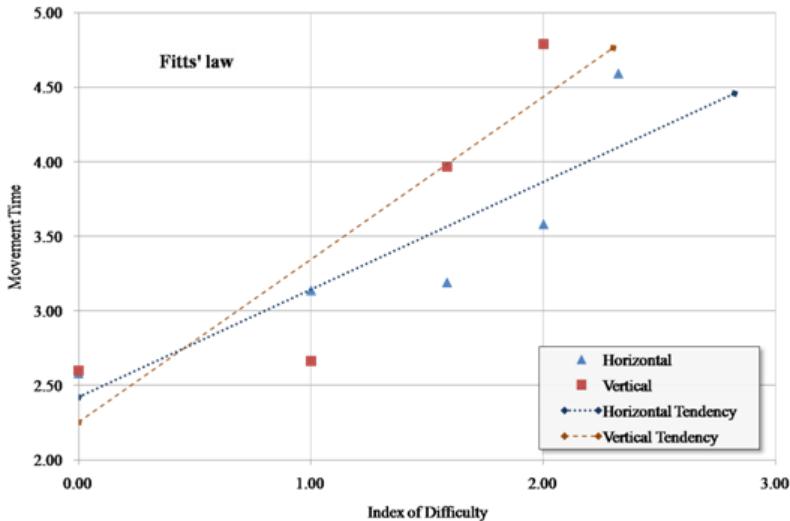


Fig. 8. The average movement time (MT) plotted against Fitts' index of difficulty (ID)

6 Conclusion

Interaction during motion will become a very important issue in mobile interfaces. Increasing number of people are interacting during motion, e.g. watching TV/video, talking on the phone, playing games. Provisions are needed to ensure people can carry out these multiple tasks concurrently e.g. by leveraging on different modalities.

In this paper, we presented a way to interact with a mobile device without looking at it and using hands that is by head motion only, in motion. With our proposed interface using the head motion and gaze, it turns out 4x2 or 3x3 grid menu is the most effective configuration (near 100% recognition, low std. and relatively more menu items) considering the number available menu items vs. selection time and

associated error rate. Between the horizontal and vertical, the horizontal head movement was more stable with longer operational range (and thus more menu items covered).

In particular, non-visual interface can also be used for visually impaired users. In addition the findings can be applied to designing augmented reality interfaces that uses head motion or gaze. This study is only the first step in the investigation of non-visual mobile interfaces. Also equally required is more stable sensing for filtering out true intentional head/gaze motion from normal moving activities (e.g. navigation) and precise virtual cursor tracking.

Acknowledgement. This research was supported in part by the Strategic Technology Lab. Program (Multimodal Entertainment Platform area) and the Core Industrial Tech. Development Program (Digital Textile based Around Body Computing area) of the Korea Ministry of Knowledge Economy (MKE).

References

1. F. Berard, "The Perceptual Window: Head Motion as a new Input Stream", IFIP Conference on Human-Computer Interaction, 238–244 (1999).
2. S. Brewster, J. Lumsden, M. Bell, M. Hall, and S. Tasker, "Multimodal 'Eyes-Free' Interaction Techniques for Wearable Devices", Proc. of ACM CHI, Florida, 463-480 (2003).
3. Crossan, M. McGill, S. Brewster and R. Murray-Smith, "Head Tilting for Interaction in Mobile Contexts", Proc. of the 11th Intl. Conf. on Human-Computer Interaction with Mobile Devices and Services, MobileHCI, Bonn, Germany (2008).
4. Crossan, J. Williamson, S. Brewster, and R. Murray-Smith, "Wrist Rotation for Interaction in Mobile Contexts", Proc. of Mobile HCI, Amsterdam, Netherlands, (2008).
5. P. Eslambolchilar, and R. Murray-Smith, "Control Centric Approach in Designing Scrolling and Zooming User Interfaces", International Journal of Human-Computer Studies, 66(12), 838-856 (2008).
6. P. Fitts, "The Information Capacity of the Human Motor System in Controlling the Amplitude of Movement", Journal of Experimental Psychology, 47, 381-391 (1954).
7. E. Foxlin, "Inertial Head-Tracker Sensor Fusion by a Complementary Separate-Bias Kalman Filter", Proc. of IEEE Virtual Reality Annual International Symposium, 184-196 (1996).
8. K. Hinckley, J. Pierce, and E. Horvitz, "Sensing Techniques for Mobile Interaction", Proc. of UIST, 91-100 (2000).
9. E. LoPresti, D. Brienza, J. Angelo, L. Gilbertson, and J. Sakai, "Neck Range of Motion and Use of Computer Head Controls", Proc. Intl. ACM Conference on Assistive Technologies, Arlington, Virginia, 121-128 (2000).
10. MacKenzie, "A Note on the Information-Theoretic Basis for Fitts' Law", Journal of Motor Behavior, 21, 323-330 (1989).
11. G. Miller, "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information," Psychological Review, 63(2), 343-355 (1956).
12. Oakley and M. O'Modhrain, "Tilt to Scroll: Evaluating a Motion based Vibrotactile Mobile Interface", Proc. of World Haptics, 40-49 (2005).
13. Oakley and J. Park, "A Motion-based Marking Menu System", Extended Abstracts of ACM CHI, San Jose (2007).

14. Rekimoto, "Tilting Operations for Small Screen Interfaces", Proc. of UIST, 167-168 (1996).
15. G. Robert, C. Gregg, "A Method for Evaluating Head-Controlled Computer Input Devices Using Fitts' Law", Human Factors, 32(4), 423-438 (1990).
16. Welford, "Fundamentals of Skill", London: Methuen, (1968).
17. Williamson, R. Murray-Smith, and S. Hughes, "Shoogle: Multimodal Excitatory Interaction on Mobile Devices", Proc. of ACM CHI, San Jose (2007).
18. S. You and U. Neumann, "Fusion of Vision and Gyro Tracking for Robust Augmented Reality Registration", Proc. of IEEE Conference on Virtual Reality, Japan, 71-78 (2001).

Eye-Movement-Based Instantaneous Cognition Model for Non-verbal Smooth Closed Figures

Yuzo Takahashi and Shoko Koshi

Ergonomics Laboratory, Graduate School of Information Sciences,
Hiroshima City University, Japan
3-4-1, Ozuka-higahi, Asaminami-ku, Hiroshima City, Japan
y-taka@hiroshima-cu.ac.jp

Abstract. This study attempts to perform a comprehensive investigation of non-verbal instantaneous cognition of images through the “same-different” judgment paradigm using non-verbal smooth closed figures, which are difficult to memorize verbally, as materials for encoding experiments. The results suggested that the instantaneous cognition of non-verbal smooth closed figures is influenced by the contours’ features (number of convex parts) and inter-stimulus intervals. In addition, the results of percent correct recognitions suggested that the accuracy of the “same-different” judgment may be influenced by the differences between the points being gazed when memorizing and recognizing and factors involved in the visual search process when recognizing. The results may have implications for the interaction design guideline about some instruments for visualizing a system state.

Keywords: non-verbal information, cognition model, eye movements, same-different judgment paradigm.

1 Introduction

The data and images demonstrating that verbalizing is difficult exist in a factory environment. These images are level meters changing over time combined with several warning lights. It is thought that these signs could not be instantaneously represented by language. Therefore, to achieve instantaneous judgment, several interfaces in factories have to support the intuitive images.

In general, familiar information is easy to memorize verbally, but a new image or an image changing over time is difficult to memorize verbally because of workers’ ability to perceive instantaneously [1]. Human working memory capacity was examined using unstructured information, and the results suggested that the capacity to memorize unstructured information was between 3 to 5 data set sizes [2] [3] [4].

To quickly and accurately recognize different information presented successively, it is necessary to establish an information presentation method that stimulates human reactions without language (verbalized information) intervention.

Some studies on non-verbal smooth closed figures suggested that the enhancing factors for memorizing this type of figures were influenced by (1) distortions of convex parts, (2) psychological similarity [5], (3) physical complexity (ex: $Perimeter\sqrt{Area}$),

(4) the number of convex parts, (5) the length of individual convex parts [6], (6) presentation and retention time [7], (7) the Fourier descriptors [8] [9], (8) spatial frequency characteristics [10], and (9) the number of figures presented simultaneously [11].

As part of a study on Ecological Interface Design in Human-Computer Interaction, the present study attempts to conduct a comprehensive research on the non-verbal instantaneous cognition of smooth closed figures through the “same-different” judgment paradigm using non-verbal smooth closed figures, which are difficult to memorize verbally, as materials for encoding experiments.

2 Method

2.1 Subjects

Ten subjects (20–22 years; 21.6 ± 0.7 years) were selected among male university students with no significant eyesight problems. Six had their vision corrected, three by wearing glasses and three by wearing contact lenses.

2.2 Apparatus

Smooth closed figures were presented on a screen using an LCD projector. The subjects' eye movements were determined by measuring the sightline of the dominant eye using EMR-NL8 (nac Image Technology) at 60 Hz. The heights of the work table, seat, and chin support were adjusted arbitrarily for each subject. The luminous intensity of the screen background was set as 230.0 cd/m^2 and the luminous intensity of contour line was set as 80.0 cd/m^2 . Our experimental setting is shown in Fig. 1.

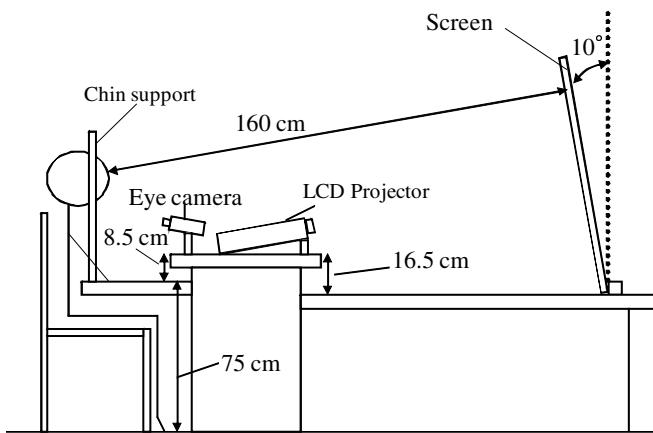


Fig. 1. Experimental setting for measuring subjects' eye movement

2.3 Procedure

After the subject observed the gazing point continuously for 2000 ms, a non-verbal smooth closed figure to be memorized was presented for 250 ms, followed by a

masking pattern, which was also presented for 250 ms. The inter-stimulus intervals (ISI) of 500, 1000, and 4000 ms were used. A recognition figure was presented after the preset ISI, and the subject was asked to discriminate the recognition figure from the memorized figure (“same-different” judgment paradigm [5] [6] [7]). A time limit of 5000 ms was enforced.

2.4 Encoding Stimuli

Based on the results of past studies [5] [6] [7] demonstrating that the performance of complex free-form figure recognition diminishes as the number of convex parts of the memorized figure increases, the number of convex parts of memorized figures was set as 5, 7, and 9 [10]. Figures as encoding stimulus were generated by the formula (1) [5] [10] and the formula (2) [5].

$$r(\theta) = A_0 + \frac{1}{3} \sum_{i=1}^5 A_i \cos(f_i \theta - \alpha_i) \quad (0 \leq \theta \leq 2\pi) \quad (1)$$

A : Amplitude, f : frequency, α : phase

$$(x(\theta), y(\theta)) = r(\theta)(\cos(\theta), \sin(\theta)) \quad (0 \leq \theta \leq 2\pi) \quad (2)$$

Influences of both local features and general features have been suggested as memory characteristics of non-verbal smooth closed figure’s contour shapes. Therefore, the present study uses both types of contour shapes, those for which the influence of general features has been suggested and those for which the influence of local features has been suggested. In addition, the relationships between the memorized figure and recognition figure were defined as follows: 1) the recognition figure being the same as the memorized figure; 2) the recognition contour intensifying the low or high frequency component of the presented figure by 3.0 or 6.0% [10]; and 3) the differences between the areas of projections and/or indentations are 3.7, 6.3, and 9.0% [11]. Figure 2 shows examples of non-verbal smooth closed figures.

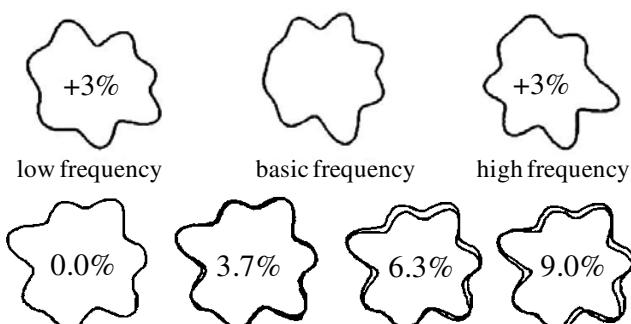


Fig. 2. Examples of non-verbal smooth closed figures

2.5 Evaluation Indexes

For recognition performance, the ratio of correct answers (percent correct recognitions) and reaction time were determined. In addition, the total length of path in the sightline (the eye movements) when memorizing and recognizing contour shapes, the area being gazed, the number of features in the perimeter of the area being gazed when memorizing (see Fig. 3.), and the number of eye fixation points and fixation time during recognition were determined from the sightline data.

3 Results

3.1 Recognition Performance

A two-factor analysis of variance was performed for percent correct recognitions (Fig. 4.) and reaction time (Fig. 5.) using the number of convex parts and ISI as factors. Significant main effects were observed for both factors in percent correct recognitions (number of convex parts: $F(2,18)=6.12, p < 0.05$; ISI: $F(2,18)=10.38, p < 0.001$). For reaction time, a significant main effect was observed only for ISI ($F(2,18)=14.59, p < 0.001$).

3.2 Eye Movements Performance

A two-factor analysis of variance was performed for the number of fixation points (Fig. 6.) and the fixation time (Fig. 7.) in the recognition phase using the number of convex parts and ISI as factors. Significant main effects were observed for the number of fixation points (number of convex parts: $F(2,18)=3.82, p < 0.05$; ISI: $F(2,18)=11.73, p < 0.001$). And a significant main effect was observed for the fixation time (ISI: $F(2,18)=5.69, p < 0.05$). To evaluate the effects of area being gazed when memorizing, a single-factor analysis of variance was performed for the area being gazed using the number of convex parts (Fig. 8.). As a result, a significant effect was observed for the area being gazed ($F(2,18)=6.42, p < 0.01$). Subsequently, to evaluate the effects on recognition performance due to the number of features of memorized figures, a single-factor analysis of variance was performed for the number of features on the smooth closed figure using the number of convex parts (Fig. 9.). A significant effect was observed for the number of features ($F(2,18)=58.62, p < 0.001$).

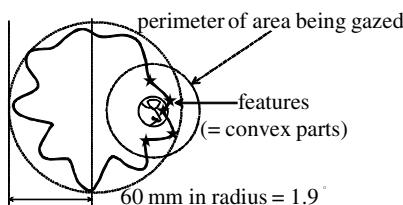


Fig. 3. Schema of features in the perimeter of the area being gazed

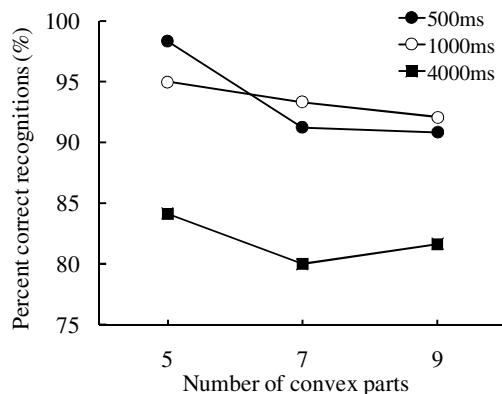


Fig. 4. Percent correct recognitions as a function of the number of convex parts with ISI

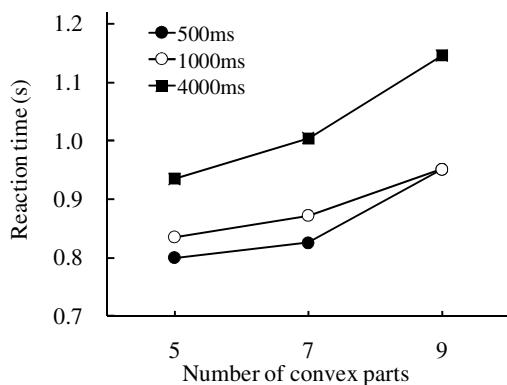


Fig. 5. Reaction times as a function of the number of convex parts with ISI

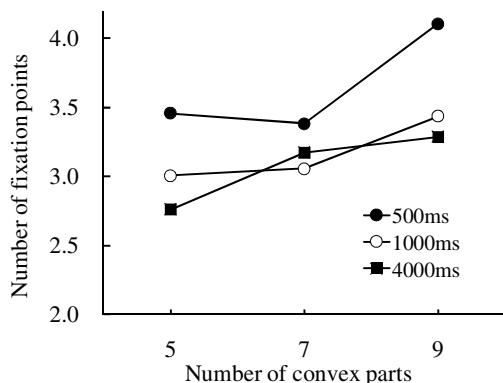


Fig. 6. Number of fixation points as a function of the number of convex parts with ISI

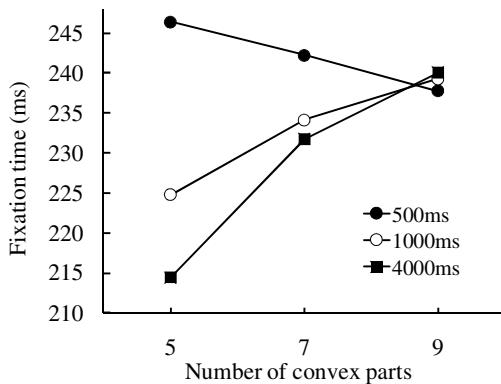


Fig. 7. Fixation time as a function of the number of convex parts with ISI

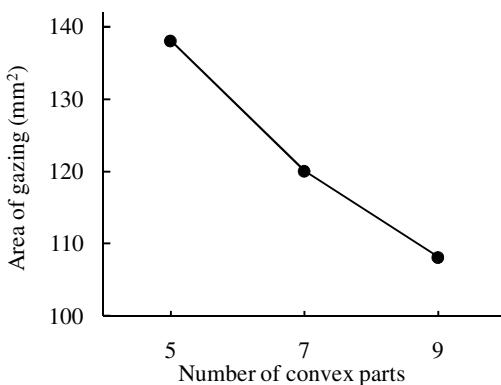


Fig. 8. Area being gazed around the contour as a function of the number of convex parts

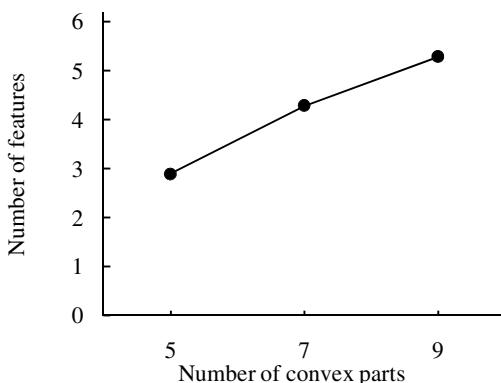


Fig. 9. Number of features around the contour as a function of the number of convex parts

3.3 Development of Instantaneous Cognition Model

To evaluate the process of instantaneous cognition and recognition for non-verbal smooth closed figures, correlations between various indices used in the experiment were calculated. After eliminating significant partial correlations, a correlation diagram was produced. To generalize the relationships between evaluation indices, a principal component analysis was performed using the indices used to produce the correlation diagram and four major components (focus of attention, fixation points, points from which subjects' began searching, and memorized features) were extracted.

Because the cognitive process of non-verbal smooth closed figures includes a cognitive function that cannot be measured objectively, a maximum-likelihood method was used for a factor analysis and the three latent variables were extracted (Features of contour shapes; F.CS, Starting point of gaze; SPG, and Visual search process; VSP), each of which has a causal relationship with a different indicator. Figure 10 shows an instantaneous cognition model in which the sequential order of cognition-recognition processes, indices, principal components, and cause-and-effect relationships between factors are considered after inserting the extracted factors.

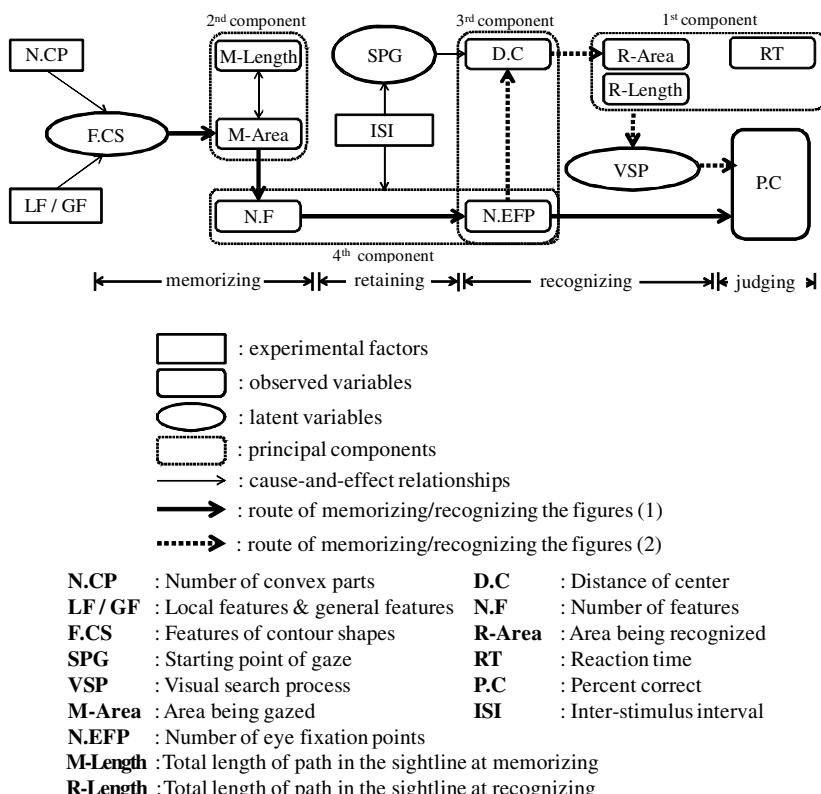


Fig. 10. Instantaneous Cognition Model

4 Conclusions

The present study evaluated the characteristics of instantaneous cognition of non-verbal smooth closed figures from the perspectives of eye movement when memorizing/recognizing the figures and recognition performance through the “same-different” judgment paradigm using non-verbal smooth closed figures. The results suggested that the instantaneous cognition of non-verbal smooth closed figures is influenced by the contours’ features (number of convex parts) and ISI. In addition, the percent correct recognitions results suggested that the accuracy of “same-different” judgment may be influenced by differences between the points being gazed when memorizing and recognizing and the factors affecting the visual search process when recognizing. Another factor suggested as having a significant effect on “same-different” judgment accuracy is the number of features projected on the retina when memorizing contour shapes. In conclusion, the results of this study may have implications for interaction design guidelines about some instruments for visualizing a system state, particularly developing the supporting devices in work requiring precise detailed observation (factories and other environments requiring quality assurance, such as multi-attribute operations). However, cognition differences may exist at different ages because of depression of visual acuity. In further study, such comparisons across several ages are needs to design the visual instruments requiring the instantaneous cognition.

Acknowledgments. This work was supported by MEXT Grant-in-Aid for Young Scientists (B) (21700142).

References

1. Becker, M.W., Pashler, H., Anstis, S.M.: The role of iconic memory in change-detection tasks. *Perception* 29, 273–286 (2000)
2. Cowan, N.: Capacity Limits for Unstructured Materials. In: *Working Memory Capacity*, pp. 105–137. Psychology Press, San Diego (2005)
3. Fisher, D.L.: Central capacity limits in consistent mapping, visual search tasks: four channels or more? *Cognitive Psychology* 16, 449–484 (1984)
4. Luck, S.J., Vogel, E.K.: The capacity of visual working memory for features and conjunctions. *Nature* 390(20), 279–281 (1997)
5. Sakai, K., Inui, T.: Properties of distortion of the figures in visual short-term memory. *The Japanese Journal of Psychology* 70(3), 211–219 (1999)
6. Sakai, K., Inui, T.: Psychological complexity of the figures and properties of retention in visual short-term memory. *The Japanese Journal of Psychology* 71(1), 9–16 (2000)
7. Sakai, K., Inui, T.: Effect of figural complexity and target-distracter similarity on decay rate in short-term visual memory. *The Japanese Journal of Psychology* 72(1), 36–43 (2001)
8. Endo, N., et al.: Perceptual judgement of novel contour shapes and hierarchical descriptions of geometrical properties. *The Japanese Journal of Psychology* 74(4), 346–353 (2003)

9. Zahn, C.T., Roskies, R.Z.: Fourier descriptors for plane closed curves. *IEEE Transactions on Computers* C-21, 269–281 (1972)
10. Takahashi, Y.: Effects of Smooth closed Figure's Spatial Frequency on Correct Recognition Task. *The Japanese Journal of Ergonomics* 41(supplement), 250–251 (2005)
11. Takahashi, Y.: Effects of the number of visually presented smooth closed figures on correct recognition task. *The Japanese Journal of Ergonomics* 40(supplement), 476–477 (2004)

Part III

Voice, Natural Language and Dialogue

VOSS -A Voice Operated Suite for the Barbadian Vernacular

David Byer and Colin Depradine

University of the West Indies,
Cave Hill, St Michael, Barbados
david.byer@gmail.com,
colin.depradine@cavehill.uwi.edu

Abstract. Mobile devices are rapidly becoming the default communication device of choice. The rapid advances being experienced in this area has resulted in mobile devices undertaking many of the tasks once restricted to desktop computers. One key area is that of voice recognition and synthesis. Advances in this area have produced new voice-based applications such as visual voice mail and voice activated search. The rise in popularity of these types of applications has resulted in the incorporation of a variety of major languages, ensuring a more global use of the technology.

Keywords: Interfaces, mobile, Java, phone, Android, voice, speech, Windows Phone 7.

1 Introduction

This paper presents VOSS a voice operated software suite for the Barbadian vernacular. Its primary function is to provide tools for translating the local indigenous grammar to Standard English. Barbados is a small Caribbean island, located to the east of the chain of islands. It is a former British colony and so English is the standard language taught in schools. However, as with most Caribbean islands, there is an internal dialect which is spoken by the local population. In fact, it is not uncommon to encounter instances where Standard English and the Barbadian dialect are being used interchangeably. This can be problematic since visitors to the island may find it difficult to understand someone using the dialect. Since tourism is one of Barbados's primary commercial sectors, the use of technology can provide a cost effective mechanism to help alleviate this problem.

2 The VOSS System

The VOSS system is a collection of integrated software and hardware components which allows both external and administrative users to access and to add or update data that is generic to the Barbadian culture and vernacular. It is a combination of web servers, databases and web based or windows forms application interfaces.

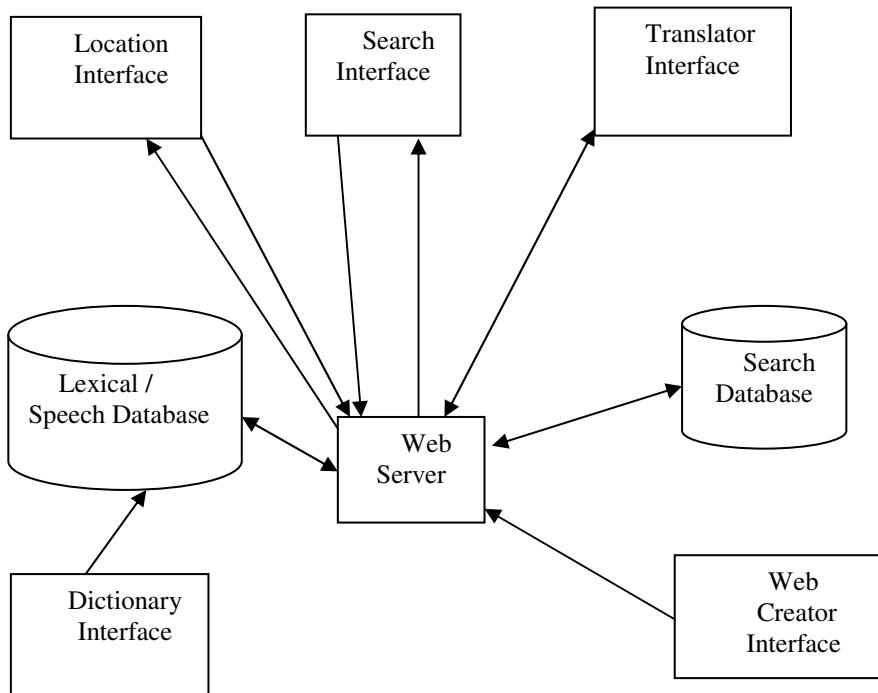


Fig. 1. Overview of the VOSS System

2.1 Dictionary Interface

This is a windows form application interface to be used by the administrators of the system for the purpose of adding words to the lexical database, which is used by the translator interface, and to the voice recognition dictionary.

This interface is password protected so that no unauthorized user can make changes to the database. The user keys the correct spelling of the word they wish to add. Then using the basic rules of syntax the various parts-of speech categories, such as noun or verb, that the word falls into are added to its properties. When that process is completed the user will pronounce the word as it should be recognized by the speech recognition engine. The system will ask the user to repeat the word three times then it will replay the word. If the user is satisfied with the pronunciation he or she can apply the word and it will be added to the dictionary. If the user is not satisfied, the process can be repeated. The interface is also used to edit or remove existing words from the database.

2.2 Web Page Creator

This interface is used to create the content on the web server. This interface does not require the user to have any knowledge of web design, instead, the content is generated using Microsoft Word.

The system uses two types of web pages. One type is used by the search interface and consists of one page documents with embedded graphics and required search keywords. The other type is used by the locator interface and is tied to the GPS and is coordinate specific.

During the creation of a general web page, the user must first create the document in Microsoft Word. The user will then select the web page creation option within the VOSS system where they will be prompted to enter a title for the page and then to browse for the appropriate Microsoft Word document. When the document is selected, the interface reads its content and creates 15 search keywords, each with a weight of importance. An HTML page is generated and uploaded to the web server along with the relevant search criteria.

However, when it comes to the data that is tied to the GPS it creates a collection of pages and groups them together with the link being the coordinate. These too are created in Microsoft Word but the difference is that they are separate pages and the user has the option to create one page for a site and upload it, stop and go to another site and then come back to it and create another page. This option allows for pages for any historical site to be changed and updated while maintaining the same file name on the server.

2.3 Location Interface

This interface is for mobile devices and is tied into the GPS system. The user uses voice commands to activate a map which shows the user's current location.

The user employs a variety of commands to change the map view to roads terrain, satellite or hybrid map topologies. Voice commands can also be used to pan left and right and to zoom in and out. This interface runs in a browser container which allows for map documents to be manipulated by the interface software.

The user can also vocally request any historical and geographical data about the location or pictures of buildings and/or landmarks in the area. This interface runs on the Microsoft Windows Phone 7 operating system, any phone using the Android operating system, Windows mobile devices and laptops.

These devices must all be GPS enabled as the software uses the GPS coordinates as a key to submit to the web server and to retrieve the pages of historical or cultural data about the particular site. The devices must also have Internet connectivity.

2.4 The Search Engine Interface

This is a standard search engine interface which accepts voice input and interprets the query string. Since many Barbadians speak both Standard English and dialect it is easy to receive sentences which are a combination of the two.

As a result it will be necessary to translate the spoken dialect query into Standard English. This is performed by an algorithm which uses the lexical database and grammar rules to translate the statement to Standard English.

Once that is completed, the search algorithm will interpret the query statement to ascertain the most accurate request to submit to the database and what keywords should be used in the search.

The engine will return various types of information about things Barbadian. The user has the option to read the data for him or herself or to instruct the interface to read it aloud for them.

2.5 The Translator Interface

This interface comes in two flavors, one for mobile devices and one for the desktop or laptop computers (PC). This is a dictation tool which is used to translate the Barbadian dialect to Standard English.

Firstly, the system takes a body of submitted text and breaks it down into sentences. Each word in the sentence is then checked against the translation tables in the lexical database. When a dialect word is found it will be replaced with a standard word or phrase.

This continues sentence by sentence until the whole document is processed. The system will then return to the beginning and use the rules of syntax and semantics to adjust the sentences to make them grammatically correct.

This interface is mainly used on a PC for the purpose of creating text documents that are in Standard English due to the fact that it is not uncommon for Barbadians to unconsciously mix dialect and English when creating documents. The user can then print or email the finished document.

The mobile side is mainly for visitors and those who do not understand Barbadian dialect. The mobile device will first record the conversation with permission of the speaker. The recording is saved as a .wav file and is transmitted to the server along with an email address. The web server passes the file to the translator where it is converted to text. At this point, the translation process carried out by the PC version is performed. When the process is completed, the translated text is sent by email to the user.

3 Platforms and Programming Languages

The system has been developed using the Microsoft C# (c-sharp) and Java programming languages. The C# language provides application programming interfaces (API) for the Microsoft speech library as well as the Microsoft .NET framework for mobile devices, specifically the Microsoft Windows Phone 7 system. Java is used for the backend processing because it has several prebuilt classes for parsing and communication over the Internet and can also be used with the Google Android mobile operating system.

3.1 Windows Phone 7

Windows Phone 7 is a mobile operating system which was developed by the Microsoft Corporation. This operating system uses a design language called Metro. Metro is used to integrate the operating system with other services provided by Microsoft and third party providers. This operating system supports two popular programming platforms which are Microsoft Silverlight and XMA [10].

Silverlight is a development platform which is used by software developers for creating interactive multimedia interfaces which can be either online or offline. Silverlight can be used for Web, desktop, and mobile applications. [8]

XNA is a runtime environment developed by Microsoft for developers who create games for its gaming platforms. It supports 3D graphics and is mostly used for developing games for the Xbox but it can be used for the PC and for the Microsoft audio player, the Zune HD [8].

3.2 Android

Android is a mobile operating system that was developed by a company called Android Inc. This company was acquired by Google in 2005. Android is open source and is based upon a modified version of the Linux kernel, [1].

The main language used for development on Android is the Java language. The Java programmer develops applications for controlling the device by using Google-developed Java libraries. Some of the other languages than can be used are Python and Ruby. [1].

4 Methodology and Testing

A modular approach is being used for the development of the VOSS system. As each module is completed it is tested before moving on to the next one. Figure 2 below shows the development flowchart for the system.

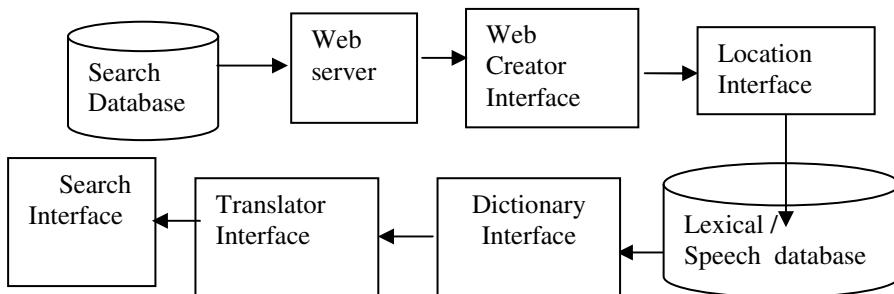


Fig. 2. Development Flow Diagram

4.1 Development Progress

The system is currently 40% completed. The operational components are the web server, the web page creator and the location interface, which are 100% completed. The ones currently being worked on are the dictionary interface, the translator interface and the search engine interface, which are roughly 50% completed.

The system was tested by creating web-based data of about ten locations in Barbados using the web creator interface and then going to the ten predetermined areas and using the locator interface to see if the map and the data about the area would be available. It was tested using a laptop with a GPS locator and a Smartphone with the Android operating system.

4.2 Results of Testing

In examining the web server database it was found that the algorithm which creates the keywords for the automatically generated web pages was only 85% accurate with some of the pages being mislabeled.

In testing the locator interface on the laptop there was an 80% accuracy rate for the GPS software that was programmed into the interface and at least in one case, the location information provided was completely incorrect. However, for the sites that were correct the system did display the correct map location and the accompanying web pages that gave the historical and or cultural details about the area.

The navigation of the map and of the web pages is currently manual as the voice component is not fully operational.

Presently test are being performed on the translator and it is currently correctly parsing through the sentences and is replacing the dialect with Standard English words or phrases but the algorithm for making the sentence grammatically correct is still in development.

References

1. Android Operating System, <http://en.wikipedia.org/wiki/Android>
2. Bangalore, T.M.: Qme!: A Speech-based Question-Answering system on Mobile Devices. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pp. 55–63 (2010)
3. Hakulinen, M.T.: Agent-based Adaptive Interaction and Dialogue Management Architecture for Speech Applications. Speech and Dialogue. In: Proceedings of the Fourth International Conference, pp. 357–364 (2001)
4. Karl, L.P.: Speech versus mouse commands for word processing applications: An empirical evaluation. Int. J. Man-Mach. Stud. 39(4), 667–687 (1993)
5. Mathias Creutz, M.C.: Web augmentation of language models for continuous speech recognition of SMS text messages. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, pp. 157–165 (2009)
6. Michael, S., David, Z., Hu, M., Prasad, S.: Managing multiple speech-enabled applications in a mobile handheld device. International Journal of Pervasive Computing and Communications 5(3), 332–359 (2009)
7. Pakucs, B.: Butler: A Universal Speech Interface for Mobile Environments. In: Brewster, S., Dunlop, M.D. (eds.) Mobile HCI 2004. LNCS, vol. 3160, pp. 399–403. Springer, Heidelberg (2004)
8. Petzold, C.: Programming Windows Phone 7. Microsoft Press, Redmond (2010)
9. Shneiderman, B.: The Limits of Speech Recognition. Communications of the ACM 43(9) (2000)
10. Windows Phone 7, http://en.wikipedia.org/wiki/Windows_Phone_7

New Techniques for Merging Text Versions

Darius Dadgari and Wolfgang Stuerzlinger

Dept. of Computer Science & Engineering

York University, Toronto, Canada

<http://www.cse.yorku.ca/~wolfgang>

Abstract. Versioning helps users to keep track of different sets of edits on a document. Version merging methods enable users to determine which parts of which version they wish to include in the next or final version. We explored several existing and two new methods (highlighting and overlay) in single and multiple window settings. We present the results of our quantitative user studies, which show that the new highlighting and overlay techniques are preferred for version merging tasks. The results suggest that the most useful methods are those which clearly and easily present information that is likely important to the user, while simultaneously hiding less important information. Also, multi window version merging is preferred over single window merging.

Keywords: Versioning, version merging, document differentiation, text editing, text manipulation.

1 Introduction

Versioning allows users to compare two (or more) versions of a single document and to create a new version, which presumably synthesizes the “best” features of the old ones. This is a common activity when jointly writing an essay or conference paper, where each participant may modify an original document in parallel to improve it. Then is a necessary for the task-leader to compare and merge text from these multiple versions towards the common goal of creating a final version. We focus on (English) text documents and do not evaluate methods for source code, alphanumeric strings, highly structured poetry, or other forms of writing. Moreover, we ignore formatting, layout and other meta-data.

However, versioning and version merging is not used universally today, due to the perceived difficulty to use, or the users’ lack of knowledge of the tools. Even though text versioning systems have been available for more than a decade, there is no published comparative examination of version merging methods. Moreover, multiple window versioning is rarely used, too. Most common version merging methods use a single window and show information as necessary within that single window. Multiple window version merging uses two or more separate windows, where each uses some (potentially different) visualization method to differentiate between versions. This reduces visual clutter.

The primary objective of this paper is to explore a variety of methods for text version differentiation and merging. More precisely, we investigate techniques that enable users to select the desired parts from among a set of different versions, i.e.

version merging. For this task, we present new techniques for both single window and multiple window versioning. Moreover, we validate these techniques with user studies in comparison to other published and commercially available techniques.

We examine whether version merging with multiple windows is superior to single-window when examining many documents, and present new interaction methods unique to multiple window text version merging.

2 Related Work

A fundamental operation in text editing is to compare two versions and to output which and where data has been added, changed, and deleted. Classical command line tools, such as `diff` [7], compute and output this list directly, thus providing a method to determine differences between documents. The output format of `diff` is very terse. This does not allow end-users to easily see how those changes relate to the context within the document and we assume that this is the main reason why non-technical people rarely use this tool. Providing the same information with overview exposes versioning to a much larger audience. `xediff` [15] and similar graphical differencing utilities, show versions side-by-side in a scrollable display and by using colored backgrounds to highlight changes. Such tools target highly structured text, such as source code, and support often only the direct merging of code changes.

Microsoft Word 2000, and similarly WordPerfect 12, provides a Track Changes option, which denotes insertions by underlining added text and visualizes deletions by applying strikethrough to the deleted text. In these applications, versioning is often referred to as “Revisions” or “Track Changes”. Insertions and deletions are colored based on which user made the change. This method is known as ‘redlining’ or ‘strikethrough’, and is called in this paper ‘*Strikeout/Underline*’.

Word 2003 utilizes balloons containing the deleted text as opposed to showing strikethrough text. This is called the *Balloon* method in this paper, which can lead to an overwhelming number of annotations. Each individual portion of changed text may be accepted or rejected relative to the previous version individually, rather than all at once. Also, a reviewing pane is provided which shows all changes in a linear top-to-bottom order. But this pane uses a lot of space even for minor changes. Finally, the variety of available options for revising can be confusing to end users and has led to the accidental public release of sensitive document data on multiple occasions. Third party plug-in software, such as iRedline [8], aim to fix some of these issues.

Word 2007 adds a one-click methods to view only comments and other formatting in balloons inline within the document. Another significant addition is the Compare Documents feature that, similar WinMerge [14], takes two documents and visualizes their similarities and differences. One issue is that four separate frames with different views of the information are shown at once, together with the title bar and other menus, which makes each pane relatively small.

Research on collaborative writing is also related to our efforts. Several studies investigated collaborative writing, with a wide variety of results. Some found that users rarely collaborate and only 24% of users write together [1]. Yet, in other studies, 87% of all documents were produced by at least two authors [3]. In cases where collaborative writing is used, one study found that the favorite function of users was

the one that allowed them to scan through the revisions made by others [12]. Users prefer version control functions as well. We speculate that such findings apply also to recently introduced collaborative writing systems, such as Google Documents.

Although source code versioning is outside of the focus of this paper, the topic is still related to general text versioning. We point the reader to the “Compare with Local History” feature of Eclipse, WinMerge [14] in combination with xdocdiff [16], the Concurrent Version System (CVS), and Subversion (SVN).

In summary, and while there are several existing implementations of text versioning methods, their usability has not been formally evaluated.

3 Version Differentiation Techniques

First, we discuss existing and two new version differentiation methods, all targeted at version merging. Later, we will evaluate these techniques with user studies. As discussed above, the most popular current document version merging techniques focus on showing all modifications to the text in a single window, which can negate the potential usefulness of multiple window techniques.

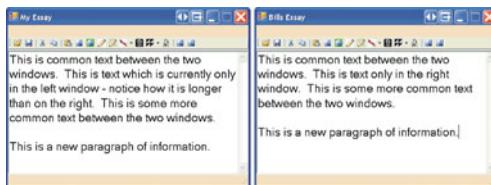


Fig. 1. Base condition: side-by-side view of two different versions of a text document

This paper examines multiple version differentiation methods that enable users to identify and compare multiple versions of a text document on-screen at the same time. Users then can select which portions of text they wish to retain in the result. This last part is most frequently implemented via appropriate GUI widgets. The cycle repeats until completion of the document. Here we describe the version differentiations methods we investigated. The base condition for our work is a side-by-side view of two versions. See Figure 1.

3.1 Highlighting Method

To differentiate text modified relative to a previous version, we present a new *Highlighting* method, which uses highlighting for changes. This focuses attention to some objects and/or diverts from others [6] and provides a simple visualization of what has and what has not been modified relative to a previous version. Considering the ubiquitous use of highlighting and highlighter markers in everyday life, this is an intuitive affordance. Our implementation highlights inserted text with a (user configurable) soft-palette background color, by default light yellow. Assuming black text on a white background, removed text is shown with a black background, which hides deletions from the user. This highlight method focuses the user on items that

likely require user attention, while diverting attention from those parts that may not require consideration. Although user attention is diverted from the deleted text, users can still easily see that text. For that, he or she can hover the cursor over the hidden text, which will turn the text color within the deleted text section white, while keeping the background color black. The user can lock or unlock this state by left-clicking on the text. This provides a simple way to view deleted text, without the added clutter of viewing all deleted text when it is not necessary. See Figure 2.

Highlighting is used to visualize differences in a variety of work, usually by highlighting important items to focus attention [10]. But, it is also recognized that too much color differentiation confuses users [17]. Thus, we believe that using too many colors for text highlighting are not positive, e.g., when the colors relate to several authors with edits that have to be reconciled. The *Highlighting* differentiation method is similar to the methods used in systems such as WinMerge, save for the addition of whitespace and the inability to hide information not in both documents. Google Documents also recently added highlighting in their revision feature.

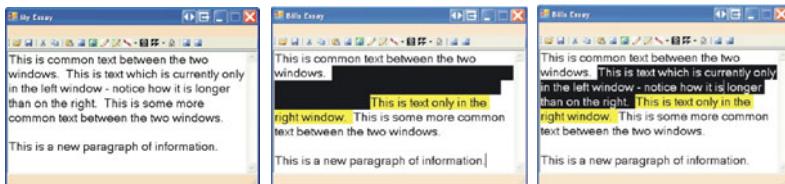


Fig. 2. Highlighting Method. Left – Original version. Middle – changed version with blackened text. Right – same version during mouse hover.

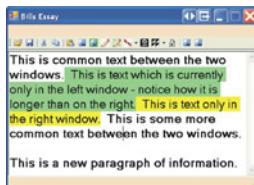


Fig. 3. The two windows of the Base Conditions overlaid with the Overlay method

3.2 Overlay Method

Transparency or translucency enables users to understand directly how two documents correlate by layering them, similar to using transparencies to overlay information onto a page of text. Based on this idea, we present the new *Overlay* method, where users can make windows transparent via alpha blending [13]. Users can then drag one window on top of another, which snaps the top window to the bottom window and locks it in place. In this state, the toolbar, menu bars and status bar of the top window are shown normally, but the text area background is made transparent, and the text position is synchronized to the uppermost paragraph in the underlying window. Text that is identical in both versions is shown shadowed and

bold with a white background. Text that is only included in the top layer is highlighted in light yellow and text only in the bottom layer in light green. In order to prevent the text from becoming garbled due to different spacing, appropriate white space is automatically added. This maintains full visibility of the data. See Figure 3. Dragging the top window off the bottom one restores the normal appearance of both windows. This process gives the user the full illusion of a translucency effect. We note that this process also bypasses the inherent difficulties in reading blended text, which can severely affect readability [5].

3.3 Strikeout / Underline Method

The *Strikeout/Underline* method, also known as ‘redlining’, mimics the Word Track Changes functionality and is similar to the *Highlighting* method. However, insertions are shown in red and are underlined, whereas deletions are drawn red and have a strikethrough effect applied. Different colors are used to differentiate between different users’ modifications. Users can hover their mouse over the text for a balloon that displays the author, date, time and type of change, see Figure 4. This method was extensively used prior to Word 2003 and can scale to multi-document merging.

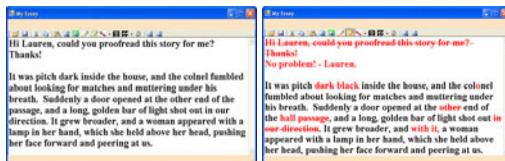


Fig. 4. Strikeout/Underline Method. Left – Text prior to edits. Right – Text after edits.

A drawback of this method is that deleted text is shown with the same text size as unmodified or inserted text. As the deletion itself suggests that the information will likely not be as important for the next revision this is counterintuitive. In addition, the strikethrough effect negatively impacts readability. Another drawback is that the color of deleted text is the same as for inserted text, which also impacts readability. There are numerous commercial implementations that enhance this basic method [2][9].

3.4 Balloon Method

This method mimics the Word 2003 Track Changes functionality. This is one of the most commonly used text version merging systems today due to the high usage rate of Microsoft Office. In this method, insertions are marked in red. Deletions are shown in the right margin in separate, fixed-position balloons that contain the deleted text. A dotted line and arrow point to the location the text was deleted from. Users can hover their mouse over the text or link to show a balloon that displays more information about the text change, thus allowing users to retrieve additional information as required, see Figure 5.

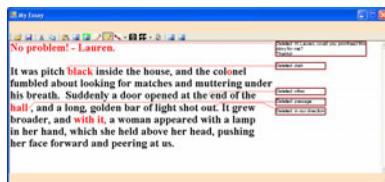


Fig. 5. Balloon Method. Same original text as in Figure 4, story after edits.

This method provides a way for users to quickly view deleted information while at the same time removing it from the main text view. It also shows where insertions occur and permits for extra information to be shown as necessary. The primary drawback is that the balloons require a significant amount of screen space. In a 1024x768 window a Word document in Print Layout at 100% zoom uses ~21% of the horizontal width (218 pixels) solely for the balloons. Another drawback is that the balloons always show the complete deleted text. Moreover, the arrows are fairly thin, and it is not always easy to see the original context of a deletion. These lines are also always visible, which creates more visual clutter. Word 2007 gives users the option to turn off the dotted lines, and they appear only if the user selects the balloon.

4 Evaluation

The goal of this study was perform an initial exploration of the topic and to compare the discussed version differentiation methods in terms of performance. We performed a 2x5 within subjects experiment to examine the ease-of-use and efficiency of the version differentiation methods. The intent was to determine which of the five compared methods were preferred. Our independent variables were window type (single and multiple window) and method used (the base condition, *Balloon*, *Strikeout/Underline*, and the new *Highlighting* and *Overlay* techniques). In the base condition we provided users with a physical copy of the older document. Our two dependant variables were error rate and task completion time. The order of the conditions and essays was counterbalanced.

In the single window conditions, one text window was shown on-screen containing a single document with annotations relating to a modification of the first text document. The size of the text workspace window was maximized to the screen. In the multiple window conditions, two equal-sized text windows filling the display were shown on the monitor. The left document window contained the original text document and the right document window contained the modified version of the left essay annotated with the given versioning method. These modifications directly related to the annotations shown in the single window condition. The sizes of the text workspace windows were identical. Furthermore, in both the single and multiple window conditions, an additional, initially minimized, target text workspace window was provided, into which participants were asked to paste information as requested.

Participants were given the following scenario: A student had written an essay a week before their assignment was due. However, the student forgot to save that essay and as such wrote it again. But, a computer-savvy friend was able to retrieve the

previously lost essay, so the student now has two similar essays that he or she can potentially hand in. Since the essay format and structure is very similar, the student can pick which portions of the essays to keep and which to throw away. Students were asked to analyze the essays paragraph-by-paragraph and copy paragraphs into the target window based on the criteria that the paragraph chosen must have no spelling errors in it and be the longer paragraph. If a paragraph with a spelling error was selected, this selection was marked as an error. If an entirely new paragraph was seen in an essay version that was not similar to the other version then participants were asked to copy that paragraph, and not doing so was also rated an error. Participants were told that other grammatical or style issues should be ignored. The length requirement is analogous to determining which part of the text is desirable and the spell-checking requirement is analogous to determining whether the specific information within a text part is the correct one. Note that copying and pasting is not representative of version merging in current text editing systems, which often have “Accept/Reject Change” functionality. However, such functionality would be in conflict with our example usage scenario that involves two different documents, which our participants found quite compelling, as almost none of them had experience with versioning. Hence and as the number of paragraphs was fixed, we decided to go with copy/paste, since the overhead for this is likely constant.

All ten used essays were approximately 1000 words in length and between 12 to 15 paragraphs for each ‘old’ and ‘new’ essay. Automatic spellchecking capabilities were not made available. Hence, participants were forced to read through all versions during the test to ensure they had no spelling errors. On purpose, the spelling errors were more obvious than commonly found in draft texts, to address the potential confound of potentially different spellchecking abilities. This ensured that participants used the investigated techniques as fully as possible.

20 participants, six female, aged 18-31, were recruited from the local University campus. All had English as their first language. Participants were paid and took 20 to 30 minutes to complete the tasks. Before the main study began, users were asked to spend five minutes to practice common operations on a test document, such as deleting lines, adding text, cutting, and pasting paragraphs. Finally, participants were given a questionnaire at the end.

4.1 Results for Single-Window Methods

A repeated measures ANOVA identified a significant main effect for task completion time and error rate overall. To simplify the explanation, we discuss the single-window method separately from the multiple-window case in the following and present only significant results of the corresponding ANOVA.

We initially examine the dependant variables of task completion time and error rate for the single window methods. Figure 6 left shows the task completion time in seconds and standard deviation for the techniques in the single-window method. The difference between the five techniques was significant ($F_{4,199}=0.7121, p<0.01$). A Tukey-Kramer test reveals that the base condition is significantly worse than all other conditions (pair wise $p<0.01$, except for *Balloon* $p<0.05$). *Balloon* is also significantly different from all others ($p<0.01$).

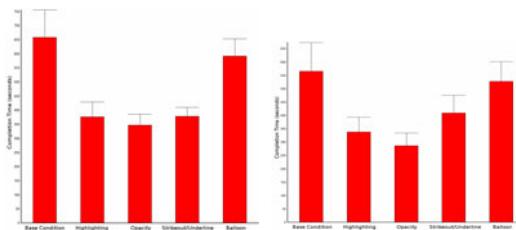


Fig. 6. Average completion time in seconds and standard deviation for Single Window (left) and Multiple Window (right) version merging methods

On a 7-point Likert scale with 7 being the best rating and 1 the poorest, the median user preference for the Base Condition was 2.0, for *Highlighting* 5.0, for *Overlay* 5.0, for *Strikeout/Underline* 5.0, and for *Balloon* 3.0.

The mean number of errors of the techniques was significantly different ($F_{4,199}=0.6293$, $p<0.01$), where an error is either a missed paragraph or a paragraph with a spelling error. The mean number of errors for the Base Condition was 3.25, for *Highlighting* 1.7, for *Overlay* 1.7, for *Strikeout/Underline* 1.8, and for *Balloon* 2.7. According to a Tukey-Kramer test, the difference between the base condition and *Highlighting*, *Overlay*, and *Strikeout/Underline* was significant. Also, the difference between *Balloon* and *Highlighting*, *Overlay*, and *Strikeout/Underline* was significant. Finally, the difference between *Strikeout/Underline* and *Overlay* was significant.

4.2 Results of Multiple-Window Methods

The right part of Figure 6 illustrates the task completion time in seconds and standard deviation for the multiple window methods. The difference between the techniques in terms of completion time was significant ($F_{4,199}=0.8715$, $p<0.01$). A Tukey-Kramer test reveals that the base condition and *Highlighting*, *Overlay*, and *Strikeout/Underline* was significant (all $p<0.01$). Also, the difference between *Balloon* and *Highlighting*, *Overlay*, and *Strikeout/Underline* was significant (all $p<0.01$, except *Strikeout/Underline* $p<0.05$).

On a 7-point Likert scale with 7 being the best rating and 1 the poorest, the median user preference for the Base Condition was 3.0, for *Highlighting* 6.0, for *Overlay* 4.0, for *Strikeout/Underline* 5.0, and for *Balloon* 4.0.

The difference between the mean number of errors between techniques was significant ($F_{4,199}=1.19$, $p<0.05$), where an error is either a missed paragraph or a paragraph with a spelling error. The mean number of errors for the Base Condition was 2.5, for *Highlighting* 1.5, for *Overlay* 1.5, for *Strikeout/Underline* 1.6, and for *Balloon* 2.05. As per a Tukey-Kramer test, the difference between the base condition and *Highlighting*, *Overlay*, and *Strikeout/Underline* was significant.

4.3 Discussion

Overall, the *Highlighting*, *Overlay*, and *Strikeout/Underline* method performed best, with all three techniques performing significantly better than the others. On average, *Overlay* performed best, but not significantly so. The results also show that both the

base condition as well as the *Balloon* method seem to be less well suited for version merging. For the *Balloon* method, we attribute this to the fact that deletions are further away from the original document, which makes *using* deleted text difficult. Also, inserting deleted text moves other text around and confuses users. The potential strength of the *Balloon* method, the ability to handle changes by many authors, was not evaluated here.

Several participants stated in the questionnaires and/or during the study that the *Highlighting* method was preferred to *Strikeout/Underline*, primarily due to the ability to hide information that was not relevant to the task. *Strikeout/Underline* also yielded annoyed comments from users regarding readability. More precisely, they found reading deleted text difficult and made comments about eyestrain here.

Comparing across the single and multiple window conditions, we find that multiple windows decrease the task completion time significantly for most methods, except for *Strikeout/Underline*. The number of errors was also reduced significantly for the base and *Balloon* condition in the multiple windows case. And the multiple window conditions are 9 to 20% faster in terms of completion time. These findings are further corroborated by the questionnaire results and comments by participants, which show a marked preference for the multiple window condition. Note that the single window condition showed proportionally more text of the document, whereas the multiple window condition showed less text per window. Even though it puts multiple windows at a relative disadvantage, the ability to view multiple annotated versions simultaneously still seems preferable to a single window. As multiple windows are almost always faster, we believe that it is overall the better choice.

Overall, we find that methods that hide likely irrelevant information (while still providing quick access on demand) and that require less user interaction are preferred and show better results. *Highlighting* is a method that exemplifies this design choice and it is likely the best choice for text comparison and version differentiation.

An interesting result side result of our work is that, with a growing number of text versions, the amount of annotations for single-window version merging becomes unwieldy. We found that expanding this information into multiple windows simplifies the version merging task.

5 Conclusion and Future Work

We examined three existing and two new methods for text versioning and merging. All are designed to help users differentiate between multiple versions of a document. All these version-merging methods were evaluated in a user study. Two different scenarios were examined: single window and multiple windows. We found that the new *Highlighting* and *Overlay* methods provided the largest benefits in completion time, error rate, and had the strongest user preference. We also found that multiple window version merging seems significantly easier for users compared to single-window version merging as the number of text versions increases. In further, currently unpublished, work on version merging on multiple monitors we also found that *Highlighting* is also the fastest and most preferred method.

For future work, we plan to address the issue that deleted sections of text in the *Highlighting* method are always visible as colored “blocks”. As such, we propose

further research into a modified *Highlighting* method in which deleted text is shown within the document as a unique, small, symbol, which can be toggled to the “normal” *Highlighting* view of the deleted text via a mouse click and back. We also plan to investigate using multiple symbols in a row to visualize the approximate length of the deletion. Also, we plan to investigate the issue of change granularity [11].

References

- [1] Couture, B., Rymer, J.: Discourse interaction between writer and supervisor: A primary collaboration in workplace writing. *Collaborative Writing in Industry: Investigations in Theory and Practice*, pp. 87–108 (1991)
- [2] DeltaView Workshare,
<http://www.workshare.com/products/wsdelview>
- [3] Ede, L., Lunsford, A.: *Singular Texts/Plural Authors: Perspectives on Collaborative Writing*. Southern Illinois University Press
- [4] Grundy, J., Hosking, J., Huh, J., Li, K.: Marama: an Eclipse Meta-toolset for Generating Multi-view Environments. In: Proc. 30th Conference on Software Engineering (2008)
- [5] Harrison, B., Kurtenbach, G., Vincente, K.: An Experimental Evaluation of Transparent User Interface Tools and Information Content. In: Proc. UIST 1995, pp. 81–90 (1995)
- [6] Horton, W.: Overcoming Chromophobia: A Guide to the Confident and Appropriate use of Color. *IEEE Transactions on Professional Communication* 34(3), 160–171 (1991)
- [7] Hunt, J.W., McIlroy, M.D.: An Algorithm for Differential File Comparison. Computing Science Technical Report #41, Bell Laboratories (1975)
- [8] iRedline: Enhanced Word Document Comparison,
<http://esqinc.com/section/products/3/iredline.html>
- [9] Litera Change-Pro, <http://www.litera.com/products/change-pro.html>
- [10] Manber, U.: The Use of Customized Emphasis in Text Visualization. In: Proceedings of the IEEE Conference on Information Visualization
- [11] Neuwirth, C., Chandhok, R., Kaufer, D., Erion, P., Morris, J., Miller, D.: Flexible Diffing in a Collaborative Writing System. In: CSCW 1992, pp. 147–154 (1992)
- [12] Noël, S., Robert, J.-M.: Empirical Study on Collaborative Writing: What do co-authors do, use and like? In: Proc. Computer Supported Cooperative Work, pp. 63–89 (2004)
- [13] Porter, T., Duff, T.: Compositing Digital Images. *Computer Graphics* 18(3), 253–259 (1984)
- [14] WinMerge differencing and merging tool, <http://winmerge.sourceforge.net>
- [15] xdiff: X-Windows file comparator and merge tool,
<http://reality.sgiweb.org/rudy/xdiff>
- [16] xdocdiff plugin for WinMerge,
<http://freemind.s57.xrea.com/xdocdiffPlugin/en/index.html>
- [17] Yang, W.: How to merge program texts. *Journal of Systems and Software* 27(2) (1994)

Modeling the Rhetoric of Human-Computer Interaction

Iris Howley and Carolyn Penstein Rosé

Carnegie Mellon University,
5000 Forbes Ave. Pittsburgh, PA USA
{ihowley,cprose}@cs.cmu.edu

Abstract. The emergence of potential new human-computer interaction styles enabled through technological advancements in artificial intelligence, machine learning, and computational linguistics makes it increasingly more important to formalize and evaluate these innovative approaches. In this position paper, we propose a multi-dimensional conversation analysis framework as a way to expose and quantify the structure of a variety of new forms of human-computer interaction. We argue that by leveraging sociolinguistic constructs referred to as authoritativeness and heteroglossia, we can expose aspects of novel interaction paradigms that must be evaluated in light of usability heuristics so that we can approach the future of human-computer interaction in a way that preserves the usability standards that have shaped the state-of-the-art that is tried and true.

Keywords: computational linguistics, dialogue analysis, usability heuristics.

1 Introduction

As computing continues to grow more ubiquitous and new and innovative types of interaction possibilities emerge in a variety of settings far away from the desktop, the question of how to formalize and evaluate these new forms of interactions becomes an increasingly important question. The relationship between man and machine will evolve in unpredictable ways, and as computers become more situated in daily life we will need to be able to learn from the new roles both entities adopt when communicating with one another. Frameworks formalizing the relationship between user and novel interface can be leveraged towards standardizing, testing, and evaluating with respect to usability heuristics. In this position paper, we begin by proposing a multi-dimensional framework¹ for analysis of human-human conversations for the purpose of better understanding these social relationships between humans and computers. We then discuss how this might eventually generalize to a wider variety of interaction styles between humans and computers.

2 Motivation

As technology develops, computers are becoming more and more autonomous, resulting in mixed-initiative interactions, a situation in which entities (both human

¹ This work was funded in part by NSF SBE 0836012, granted to the Pittsburgh Science of Learning Center.

and computer) in the interaction contribute what is most appropriate at the most appropriate time [1]. Additionally, more systems developers are adopting the approach of adjustable autonomy, where the system can automatically modify its level of independence and control in order to help the user without becoming overly intrusive [2]. Determining when the agent should pass control to the user is already an established problem within the area of adjustable autonomy research, and as such, various methods have already been established to determine the successes and failures of these systems [3]. For example, Scerri and colleagues [4] introduce a generalized transfer-of-control strategy which includes two machine learning approaches, namely C4.5 and Markov decision processes. These approaches produce control structures that are rule-based (i.e. “if the department head is not at the meeting and it is a Monday, keep control”) or constraint-based (i.e. expected quality of the agent’s decision) [4]. However, these approaches are limiting in that they can be thought of as treating interaction overly mechanistically and thus ignoring the social aspects of the interaction. Research has already established that humans behave in social ways with systems that are not intentionally social, such as in Forlizzi’s work [5] where humans establish social relationships with their non-social, domestic vacuuming robots. Once humans begin to engage socially with computational systems, they begin to adopt a different set of expectations and to orient towards these expectations in their behavior. These behaviors and expectations then lead to differences in the balance of control and responsibility than what is typical in current more typical interaction paradigms. Thus, computational systems that potentially elicit this type of engagement with humans need to not just act accurately, but act with proper consideration to their social role within the interaction.

The majority of work on these approaches happens within the sphere of human-robotic interaction, but as personal computers become more embedded and intelligent, mixed-initiative and adjustable autonomy systems will become a reality on a daily basis. This emerging reality leaves us with important questions to answer. When the computer performs actions on its own initiative, without being commanded by the user, how will we evaluate what effect that action has on the system’s ability to reduce user memory load, provide feedback, or prevent errors [6]? From a designer’s perspective, how will we formalize what happens in these new interactions so that we may evaluate how the properties of the new interactions affect usability of the system?

Jarvis: Test complete. Preparing to power down and begin diagnostics.

Stark: Yeah. Tell you what. Do a weather and ATC check.

Stark: Start listening in on ground control.

Jarvis: Sir, there are still terabytes of calculations needed before an actual flight is...

Stark: Jarvis! Sometimes you got to run before you can walk.

- “Iron Man” Film Script, 59:10.

As an example, the dialogue segment above shows a sample piece of fictional interaction between a somewhat autonomous computer (Jarvis) and its user (Stark). Traditional usability standards can still apply in such situations. For example, the first line where the user is being kept current with the status of the machine is exemplary of the system providing good feedback. Later, when Jarvis attempts to warn Stark

before he requests something exceedingly taxing on the system could be considered approaching the prevention of an error with serious consequences (i.e., system destruction). However, we see that the user ignores the computer's warning, and the computer lets him do so. Typically offering users the option of a manual override is considered a positive usability feature. In the case of a conversational interaction where negotiation over control and responsibility is more subtle than a manual override button, it's not always apparent where the transfer of control and responsibility should and does take place. Thus, despite following some basics of usability design, there is still a serious breakdown that occurs shortly after this event in the script. This suggests a possible tension between the usability heuristics as computer systems take on more intelligent roles and interactions with them become more of a dance. In order to determine what about this interaction leads to these positive and negative assessments of usability heuristics we formalize the stylistic aspects of the interaction using a conversational analysis framework.

Past work [7] has already examined how lessons from human-human interaction in the social sciences can help inform the design of ubiquitous computing systems. In a similar spirit, in this paper we propose using methods from human-human interaction in sociolinguistics to develop a methodology to help evaluate emerging interaction paradigms. We can apply past work on human-human dialogue analysis to formalize features of human-computer interactions, perhaps even where they are not explicitly conducted in natural language, in order to investigate the usability of these new systems. The process we envision for formalizing the structure of interactions and applying to evaluating the usability of systems is shown in Fig. 1.

For the past four years our research group has worked on a computational framework for the analysis of leadership in group conversations [8], [9], [10], [11]. The work we have done has mainly been applied in the area of Computer-Supported Collaborative Learning. Evidence from our work in this area has shown that leadership is an important multi-faceted construct for investigating social positioning within a group learning context. We examine the role of leadership in conversation and social positioning through three dimensions: displays of reasoning (i.e. transactivity), contribution authoritativeness (i.e., Negotiation), and contracting/expanding for additional viewpoints (i.e., Heteroglossia). While previously this framework was applied to human-human communication, as humans were the most embedded participants in a conversation, increasingly, computers are more involved as active participants in interactions.

The majority of interactions in traditional computing environments are not conducted conversationally. However, we argue that our framework provides a conceptualization that can be adapted for other purposes. For example, the Negotiation framework embodies the idea that a display of needing validation or requiring information is a sign of non-authoritativeness. On a simple level, detection of uncertainty in a user's interaction with the system could be seen as operating on this dimension, for example, timing that indicates hesitation, or perhaps a style of touch that indicates tentativeness. The Heteroglossia framework might relate more to the quantity of constraints indicated by users in the preferences. Users who make extensive use of customization options could be seen as behaving more authoritatively.

Users who are authoritative in a functional way will develop creative ways of using the potential offered by the environment. This shows an awareness of what the system offers while not abdicating control to the system. This multi-dimensional framework would suggest that in a mixed-initiative environment, we want to see users achieving their own goals while taking advantage of what the system has to offer. The triangulation enabled through an integration of the three dimensions of this framework allows for identification of users who are adopting a dysfunctional authoritative stance, and ideally, support for adopting a more functional and effective stance within the environment.

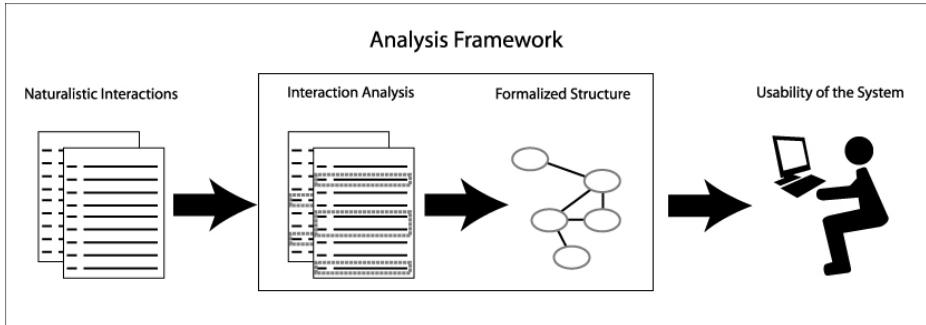


Fig. 1. Shows the process of how our analysis scheme allows us to formalize aspects of the structure of more naturalistic man-machine interactions so that we can systematically investigate how the properties of such interactions affect the usability of the system

3 The SOUFFLÉ Framework

In this paper, we focus on two dimensions of the SOUFFLÉ Framework. We consider the idea of leadership from two directions: first in terms of how authoritative a speaker presents herself, and second, in terms of how receptive a speaker is to the leadership of others. In terms of how authoritatively a speaker presents herself, we adapt two constructs from the field of systemic functional linguistics, namely Martin and Rose's Negotiation Framework [12] and Martin and White's operationalization of heteroglossia [13]. In the Negotiation Framework, authoritativeness is demonstrated by making a contribution to a discourse that is not offered as an invitation for validation from another group member. In contrast, within the Heteroglossia Framework, assertions framed in such a way as to acknowledge that others may or may not agree, are identified as heteroglossic. In representing these two dimensions separately we are able to identify those rare leaders who can present their ideas as standing on their own without denying others their own voice. Each of these dimensions presents a different take on how authoritatively a speaker is positioning himself within an interaction, which provides a rich view of leadership that allows us to see that a speaker's role adoption may not be complete, or may be dysfunctional in some way. Further details on the dimensions of our framework can be found in [11].

3.1 The Negotiation Framework

The Negotiation Framework is a measure of authoritativeness where authority is demonstrated by making a contribution to a conversation that is not meant to be validated by another group member. Past work has shown a relationship between authoritativeness and group self-efficacy within computer-supported collaborative learning tasks, speaking to the importance of authoritativeness in group functioning [14]. As part of the process of exploring this authoritativeness relationship between speakers, we have developed a coding manual so we may better identify the ebb and flow of social power within a conversation. The Negotiation Framework includes four core moves, and two secondary moves:

- K1 (Primary Knower), in which the speaker considers herself to be the primary authority on a given (expressed) piece of knowledge
- K2 (Secondary Knower), when the contributor asks for knowledge from someone of higher authority
- A1 (Primary Actor), for contributions that display that one can perform a particular action.
- A2 (Secondary Actor), when instructing someone else to do an action, allowing the other person to either perform the action or reject the request.
- ch (Challenge), in which a speaker rejects the authority of the previous speaker to make the previous move
- o (Other), which encapsulates all other moves that do not fit in the five described above

For our purposes, “Primary Knower” and “Secondary Actor” moves are considered more authoritative (with respect to social relationships), while “Secondary Knower” and “Primary Actor” moves display less authoritativeness. As such, to compute a meaningful ratio for the authoritative moves, the formula would be: $(K1 + A2) / (K1 + K2 + A1 + A2)$. It should be noted that these moves are only applicable for situations where contributions are being directed at other members of the conversation. There are several more complications to the coding process using the Negotiation Framework, a discussion of which can be found at [11]. We will explore several examples in the next section which should clarify these concepts considerably.

3.2 The Heteroglossia Framework

The Heteroglossia framework is operationalized from Martin and White’s theory of engagement [13], and here we describe it as identifying word choice that allows or restricts other possibilities and opinions. This creates a rather simple divide in possible coding terms for contributions:

- Heteroglossic-Expand (HE) phrases tend to make allowances for alternative views and opinions (such as “She *claimed* that usability heuristics are great.”)
- Heteroglossic-Contract (HC) phrases attempt to thwart other positions (such as “She *demonstrated* that usability heuristics are great.”)
- Monoglossic (M) phrases make no mention of other views and viewpoints (such as “Usability heuristics are great.”)

In some situations hedging words such as “probably” and “I think” also have a similar distancing effect as either of the Heteroglossic codes, despite the fact that the speaker may simply be unable to commit to a statement due to lack of knowledge or any other reason. Our research group has examined giving dialog agents heteroglossic, monoglossic, and neutral language in an idea generation task, and found that dialog agents with heteroglossic language result in the greatest idea generation productivity in a group task [15]. Because the level of heteroglossia an agent embodies has important effects on user productivity, we chose to focus on it in our sample analyses below.

4 Application to More Informal Interactions

User actions can be interpreted with respect to authoritativeness and heteroglossia and in this section we will go into the specifics of how looking at these actions and interactions between computers and humans can inform interaction design.

As part of our work, our research group has collected a large quantity of datasets detailing interactions between multiple humans, between a human and a computer, and between multiple humans and a computer. This collection of discussion datasets reflect what the current state of computer-human conversation interactions are currently, but what do we know about where these conversational interactions are headed? How can we use our knowledge of how the social relationships between humans and computers are changing to inform our design of interaction systems?

Here we introduce some snippets of script from the movie, “Iron Man”. We specifically selected conversational interactions between the lead character, Tony Stark (who is the Iron Man) and his home computer, Jarvis. We believe these interactions are informative as idealized interactions between man and machine, and intend to use an analysis of the script to inform future interactive systems without limiting ourselves solely to current dialogue systems. The authors would like to stress that we in no way intend to imply that the selected scenes between Jarvis and Stark from the “Iron Man” movies is indicative of future user interfaces. It is simply an appropriate example in which to show the features and flexibility of our framework for addressing the kinds of questions that arise when we consider the potentials of interactions enabled through emerging technologies in artificial intelligence and language technologies.

Table 1 is a brief interaction between Stark and his sentient computer, Jarvis, from near the beginning of the movie. We see in lines 1-4 that Stark is interacting with Jarvis in a rather authoritative way (i.e., commanding action in lines 1 and 3). These direct commands can be translated to non-dialog interaction with a system as clicks of a mouse. “Give me an exploded view” could be the equivalent of using menus to change the view of a particular system. In the film, Jarvis performs the actions commanded to him, and so we see how these commands place Stark in a considerably more authoritative role than the computer. Stark’s commands, coded within the Negotiation framework as A2, reflect a formalized version of this authority that can later be used to evaluate the interaction design.

Table 1. A short interaction between Stark and his computer from the film, “Iron Man” at time point 10:35

Line	Text	Neg.	Heterog.
1	Stark: Give me an exploded view.	A2	M
2	Jarvis: The compression in cylinder three appears to be low.	K1	HE
3	Stark: Log that.	A2	M
4	Stark: I'm gonna try again, right now.	A1	M
5	Stark: Hey, Butterfingers, come here.	A2	M
6	Stark: What's all this stuff doing on top of my desk?	K2	M

The Heteroglossia analysis also reflects the commanding nature of Stark’s requests. We see in line 2 that Jarvis allows for other possibilities with the word “appears”, rather than other options such as “is reporting” instead. In comparison to Jarvis’ heteroglossic statement, Stark’s words are entirely monoglossic commands, not allowing other options. In lines 5 and 6 Stark’s conversational style with the computer changes. He sarcastically calls the computer “Butterfingers” and appears irked at the machine as he asks the question in line 6. Stark still seems in a more authoritative stance than Jarvis with his A2 Negotiation command, but the interaction style has changed and the Negotiation codes reflect this. Now, instead of simply telling the system to perform an action, Stark asks the computer an open-ended question, denoted by the K2 Negotiation code. Line 6 shows us that the computer is potentially capable of helping its user in confusing situations.

Similarly, Table 2 shows a small portion of conversation between Stark and his conversational computer, Jarvis, from a much later point in the film. The tone has changed from the previous conversation segment yet again, with both contributors taking on mostly authoritative roles (K1 and A2 codes). In line 9, Stark has taken to adjusting his language to be less commanding. Instead of demanding “Open a new project file”, Stark states that he’d “like to open a new project file”. Similarly, in lines 11 and 12 we see in the Heteroglossia coding that Stark has opened up his statements to other possibilities. Words such as “actually” and particular phrasing like “why don’t we” show that Stark is aware of other options. This is behavior seen throughout the script being performed mostly by Stark, with Jarvis rarely using heteroglossic statements.

Unlike the previous example where Stark’s monoglossic commands can be translated into direct interface manipulations, heteroglossic statements cannot. In this situation, our framework can formalize this emerging interaction style and allow us to evaluate different approaches to handling it. We can still examine this interaction from a current usability heuristics standpoint. For example, Jarvis’ line 10, “Shall I store this on the Stark Industries Central Database?” is likely an obvious attempt at reducing the user’s cognitive load [6]. Suggesting a file-save location may save Stark from spontaneously recalling the most relevant save location from memory. Furthermore, Jarvis’ awareness of the privacy concerns related to the document in discussion can suggest putting effort towards merging security and privacy with human-computer interaction, a field further discussed in [16].

Table 2. A clip of script from the movie “Iron Man” occurring around time 54:05

Line	Text	Neg.	Heterog.
7	Stark: Jarvis, you up?	K2	M
8	Jarvis: For you, sir, always.	K1	M
9	Stark: I'd like to open a new project file, index as Mark Two.	A2	M
10	Jarvis: Shall I store this on the Stark Industries Central Database?	A2	M
11	Stark: Actually, I don't know who to trust right now.	o	HC
12	Stark: Till further notice, why don't we just keep everything on my private server?	A2	HE
13	Jarvis: Working on a secret project, are we, sir?	K2	M
14	Stark: I don't want this winding up in the wrong hands.	K1	M
15	Stark: Maybe in mine, it can actually do some good.	K1	HE

Table 3 shows another conversation between Stark and Jarvis from a later point in the movie. Jarvis’ lines in this segment are fairly typical of appropriate system feedback in an attempt to prevent errors, yet again (this time with a better ending than in the example from the introduction). The importance of the warning Jarvis communicates requires a greater level of persistence, until Stark acknowledges the warning in line 24 and essentially minimizes it. This exchange is reflected in the Negotiation coding with an A2-o repetition, with Stark once again taking an authoritative role. Stark’s commanding position is reflected in the Heteroglossia coding as well, with entirely monoglossic codes. Since this particular interaction between Stark and his computer ended successfully, we could use these codes to examine what was unique in this interaction to make it successful. Was it the A2-o pattern, or the entirely clear-cut monoglossic statements that led to the success of this interaction? Are these desirable interaction styles?

Table 3. A piece of script from the movie “Iron Man” occurring later in the movie

Line	Text	Neg.	Heterog.
16	Stark: Take me to maximum altitude.	A2	M
17	Jarvis: With only 15% power, the odds of reaching that...	ch	M
18	Stark: I know the math! Do it!	A2	M
19	Jarvis: Thirteen percent power, sir.	o	M
20	Stark: Climb!	A2	M
21	Jarvis: Eleven percent.	o	M
22	Stark: Keep going!	A2	M
23	Jarvis: Seven percent power.	o	M
24	Stark: Just leave it on the screen!	A2	M
25	Stark: Stop telling me!	A2	M

We see in these episodes a difference in footing between the user and machine that is indicated through differences in the codes assigned by the conversational analysis framework. As we then find differences in desirability between these modes of communication and collaboration between man and machine, we can use this framework to make precise what aspects of the interaction account for these differences, and thus use this analysis approach as a step towards informing design.

5 Conclusion

Our analyses from the previous section shows that our framework can formalize several of the social dimensions occurring in a more informal interaction between a human and a computer. These formalizations can then be used for evaluation of systems or to inform future designs where the relationship between human and computer is less obvious. Basing our frameworks on well-developed sociolinguistic theories that we formalize provides us with a powerful lens for challenging these new interaction techniques in the light of issues such as authority and responsibility.

In conclusion, we have shown that these frameworks originally developed for human-human discussion can be extended to human-computer dialog, inform usability for more general human-computer interaction, and it can be argued that its utility reaches even further. We may be able to use the frameworks to explore usability not just for more informal interactions and dialog systems, but possibly for even more naturalistic methods of interacting with computers, such as via gestures. While we have not, as of yet, explored how our frameworks may inform usability within gesture interactions, it seems plausible that one could detect more and less authoritative movements or more and less expansive gestures. For example, quick movements and tense body language could imply authority. The flexibility of our framework to formalizing emerging interaction systems is one of the strengths of the SOUFFLÉ framework, which allows us to expose the structure in human-computer interaction for current systems and systems yet to be invented.

References

1. Goodrich, M.A., Schultz, A.C.: Human-Robot Interaction: A Survey. In: *Foundational Trends of Human-Computer Interaction*, vol. 1(3), pp. 203–275 (January 2007)
2. Scerri, P., Pynadath, D., Tambe, M.: Adjustable Autonomy in Real-World Multi-Agent Environments. In: *Proceedings of the Fifth International Conference on Autonomous Agents (AGENTS 2001)*, pp. 300–307. ACM Press, New York (2001)
3. Fleming, M., Cohen, R.: A Utility-Based Theory of Initiative in Mixed Initiative Systems. In: *The IJCAI 2001 Workshop on Autonomy, Delegation, and Control: Interacting with Autonomous Agents* (2001)
4. Scerri, P., Pynadath, D.V., Tambe, M.: Why the Elf Acted Autonomously: Towards a Theory of Adjustable Autonomy. In: *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems: part 2 (AAMAS 2002)*. ACM Press, New York (2002)
5. Forlizzi, J.: How robotic products become social products: An ethnographic study of cleaning in the home. In: *Proceedings of HRI 2007*, pp. 129–136 (2007)

6. Nielson, J.: *Usability Engineering*. Academic Press, San Francisco (1993)
7. Bellotti, V., Back, M., Edwards, W.K., Grinter, R.E., Henderson, A., Lopes, C.: Making Sense of Sensing Systems: Five Questions for Designers and Researchers. In: Proceedings of the Conference on Human Factors in Computing Systems (CHI 2002), pp. 415–422. ACM Press, Minneapolis (2002)
8. Joshi, M., Rosé, C.P.: Using Transactivity in Conversation Summarization in Educational Dialog. In: Proceedings of the SLaTE Workshop on Speech and Language Technology in Education (2007)
9. Rosé, C.P., Wang, Y.C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., Fischer, F.: Analyzing Collaborative Learning Processes Automatically: Exploiting the Advances of Computational Linguistics in Computer-supported Collaborative Learning. *The International Journal of Computer-Supported Collaborative Learning* 3(3), 237–271 (2008) (submitted)
10. Ai, H., Sionti, M., Wang, Y.C., Rosé, C.P.: Finding Transactive Contributions in Whole Group Classroom Discussion. In: Proceedings of the International Conference of the Learning Sciences (2010)
11. Howley, I., Mayfield, E., Rosé, C.P.: Linguistic Analysis Methods for Studying Small Groups. In: Hmelo-silver, C., O'Donnell, A., Chan, C., Chin, C. (eds.) *International Handbook of Collaborative Learning*. Taylor and Francis, Inc., Abington (to appear)
12. Martin, J.R., Rose, D.: *Working with Discourse: Meaning Beyond on the Clause*. Continuum (2007)
13. Martin, J.R., White, P.R.R.: *The Language of Evaluation: Appraisal in English*. Macmillan, Palgrave (2005)
14. Howley, I., Mayfield, E., Rosé, C.P.: Missing Something? Authority in Collaborative Learning. In: Proceedings of the 9th International Conference on Computer-Supported Collaborative Learning (to appear, 2011)
15. Kumar, R., Beuth, J.L., Rosé, C.P.: Conversational Strategies that Support Idea Generation Productivity in Groups. In: Proceedings of the 9th International Conference on Computer-Supported Collaborative Learning (to appear, 2011)
16. Johnston, J., Elof, J.H.P., Labuschagne, L.: Security and Human Computer Interfaces. *Computers & Security* 22(8), 675–684 (2003)

Recommendation System Based on Interaction with Multiple Agents for Users with Vague Intention

Itaru Kuramoto¹, Atsushi Yasuda¹, Mitsuru Minakuchi², and Yoshihiro Tsujino¹

¹ Kyoto Institute of Technology, Matsugasaki, Sakyo-ku, Kyoto 606-8585, Japan

² Kyoto Sangyo University, Motoyama, Kamigamo, Kita-Ku, Kyoto 603-8555, Japan
ent@hit.is.kit.ac.jp

Abstract. We propose an agent-based recommendation system interface for users with vague intention based on interaction with multiple character agents, which are talking each other about their recommendations. This interface aims the user to make his/her intentions and/or potential opinions clear with hearing agents' conversation about recommendations. Whenever the user hits on any opinion, he/she can naturally join the conversation for getting more favorite recommendation. According to the result of experimental evaluation, the system with proposed interface can introduce more recommendations without any additional frustrations than the conventional recommendation systems with single agent.

Keywords: Vague intention, Character agent, Recommendation, Natural conversation.

1 Introduction

In the near future, there will be many recommendation systems, which recommend some items from a certain set of much information in order to support our decision making, and we will use such systems in common and frequently. In popular systems, we need to tell our intention and/or opinions to the system for getting recommendations. For example, we usually input some keywords to a book-recommendation system for purchasing interesting books. Such a system forces us to tell explicit keywords related to our needs, but we sometimes have no clear intention or opinions about our needs. In such cases, we cannot make any keywords for telling to the system, so interaction with the system stops.

In case of casual communications among three or more people like chatting, such a problem rarely occurs. If some person has no clear words to speak about a current topic in chatting, he/she can be silent because the other people continue chatting. Moreover, the silent person may make his/her opinion clear during listening to the conversation of others, because there maybe some fruitful information about the current topic, which can stimulate for his/her imagination. When he/she hits on his/her opinion and wants to talk about it, he/she can naturally go back to the conversation kept by other chatting members.

In order to solve the problems of recommendation systems about users' unclear intention, we introduce the nature of the casual communication by three or more

people. We propose an agent-based recommendation system interface which has two or more character agents who communicate with the user of the system for decision making. They talk each other if the user does not speak anything. The user can hear the conversation by the (two or more) agents, which includes some properties about some candidates of recommendation. The information can be some supports for his/her decision making. The user can rejoin to the conversation of them when he/she wants to say his/her opinion, so the user can take flexible communication to them.

2 Related Work

There are many recommendation systems with character agent interface. Moreover, there are some researches about the platform of conversational agent interface [1] so that agent-based recommendation systems will be appeared commonly in near future. However, most of such recommendation systems have only one agent [2].

Inhabited Market Place [3] is one of the agent-based presentation systems with multiple character agents collaborating with each other. In the system, the agents talk each other about an item (a car in [3]) for easy understanding about the item and for clarifying a user's unclear intention. Moreover, the user can interact to them with text input. The system is not a kind of recommendation systems, however, so it cannot change its presentation items for the user's preferences.

Sakamoto et al. proposed a competitive recommendation system with two character agents [4]. It is quite similar to our system. However, it focuses on the rationality of recommendation, so there are few discussions about interaction with agents and user. For example, it requires user's reactions ("Good" or "No good") at any ends of agents' speeches, so the user is forced to make a decision even if his/her opinion is unclear. Our study focuses on the aspect of natural interaction, so our system can avoid such situations.

Group recommender system is one of the instances of multiple character agent recommendation systems. Travel Decision Forum [5] is one of such systems. In the system, agents act a role of some members who is absent from a decision making meeting. The real members talk to the agents about the decision, and the agents tell opinions like as the absent members.

In addition, conversation between human and robot is one of the current topics [6]. It is known that the nature of human-robot interaction is different from that of human-screen agent interaction [7]. It will be valuable to know the effect of our proposed recommendation method with robot as a representation of agent.

3 Prototype

We implement a prototype system with the proposed interface (shown in Figure 1). It recommends a restaurant for lunch near Kawaramachi station, which is one of the main stations in Kyoto, Japan. There are three character agents on the system, one is a moderator agent and two are competitor agents. A user interacts with them using three opinion buttons, that is, "I agree (your recommendation)", "I do not agree", and "Give more details" (see the bottom of Figure 1).



Fig. 1. Overview of restaurant recommendation system with multiple agents

3.1 Algorithm

At first, one of two competitor agents gives his/her recommendation to the user and the other agent. In case that the user does not push any buttons, which means he/she has no opinions about the recommendation, the other competitor agent gives his/her own recommendation. If the user does not push any buttons yet, the two agents continue to talk about their recommendations.

In the conversation, one competitor agent (named Alice below) tells about a certain property of Alice's recommendation, and then the other (named Bob below) replies Alice with saying the property of Bob's recommendation. For example, Alice says "My love restaurant is near, only 3 minutes far from here by walk", and Bob replies "Good. It takes 5 minutes from here to my recommendation", and so on. The moderator agent (named Ken below) prompts them in order to keep natural and smooth conversation, and does not tell any opinions eventually.

If the user pushes "I agree" button during Alice is saying something, which means he/she decides to follow Alice's recommendation, Ken asks him/her to confirm, and then the recommendation process is finished. On the other hand, if the user pushes "I do not agree" button during Bob is speaking, Bob withdraws his opinion and recommend another new restaurant. If the user pushes "Give more details" button, the speaker (Alice or Bob) talks about another property of his/her recommendation.

In case of no user's decision and no more detail information for chat, Ken selects one of their recommendations (randomly in the prototype system) and asks the user to confirm it. If the user disagrees it, the agent who recommends it changes his/her recommendation and then return chatting. If the user does not reply again, Ken recommend the selected one and finish the recommendation process.

3.2 Implementation

The prototype system has the information of 156 restaurants which are less than 1.0 km from Kawaramachi station. The information includes the name, map, photo and properties of the restaurants such as genre, average cost, distance (by minutes' walk), and chef's recommendation menu. In this prototype system, the first recommendations by Alice and Bob are randomly selected.

The conversation is shown by text. Only one speech of a certain agent is shown at one time, so there is a 3.0-second space between the end of a certain agent's speech and the start of the next agent's speech in order that the user can catch up the conversation and tell his/her opinion by the opinion buttons.

Alice and Bob have sub-windows one for each (shown in Figure 2). It displays the name, map and photo of currently recommended restaurant by Alice or Bob. A matrix barcode on the upper-right of the window is the URL of the restaurant's website (if it has). A user can easily get the information of the recommended restaurant with his/her feature phone.



Fig. 2. Sub-window for current recommendation

4 Evaluation

4.1 Method

We asked 12 participants to evaluate the proposed system by comparing it to the conventional one shown in Figure 3. The conventional system has only one agent (named Charlie below). The system has the same interaction style (three buttons) as the proposed one.

In the conventional system, Charlie says some properties of his recommendation, and then the user decides whether he/she agrees Charlie's recommendation or not, or requests some more details for decision with three buttons.

In the evaluation, we ask the participants to use either conventional or proposed system to decide the restaurant for the day's lunch, and at the next day they use another one. We measured the number of recommended restaurants and the time of usage. After the two usages of recommendation systems, we asked participants to answer the questions below:

- Q1: Which system do you feel easy to make decision?
- Q2: Which system's recommendation do you think better?
- Q3: Which system do you feel boring?



Fig. 3. Conventional system (with single agent)

4.2 Result

Figure 4 shows the result of the average time of recommendation (a) and the average number of restaurants recommended (b). As a result, the time for interaction with the proposed system is longer than with the conventional one. However, the proposed system could introduce more recommendations to the participants than the conventional one.

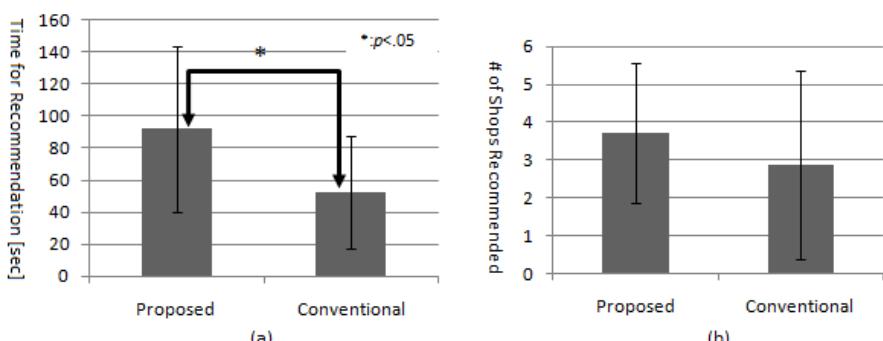


Fig. 4. The result of evaluation: (a) the average time for recommendation, (b) the average number of restaurants recommended.

Table 1. The result of questions after using recommendation system

	Proposed	Conventional	Same
Q1	3	5	4
Q2	2	5	5
Q3*	1	9	2

(*: significantly different ($p < .05$))

Table 1 shows the result of questions after the usages of recommendation systems. The result of Q3 indicates that recommendations with the proposed method give no additional frustrations even though the time for recommendation is significantly longer. It means that the system has a potential for supporting preferable decision making for users.

By contrast, the results of Q1 and Q2 indicate that the participants satisfy slightly more with the conventional system than the proposed one. In interviews after the evaluation, some participants pointed out that they preferred the conventional system because they could think deeply about the restaurant recommended. The proposed system cannot stop the conversation because of giving more information for users without explicit intention or opinions, so there was not enough time for them to make decision with deep thought. The function of reviewing conversations is needed for solving such a problem.

5 Conclusion and Future Work

We proposed an agent-based recommendation system interface which has two or more character agents who communicate with the user of the system for decision making. They talk each other about their recommendations, and the user can join and tell his/her opinions whenever he/she like, such as when his/her intention becomes clear by their conversation.

As a result of experimental evaluation, the time for recommendation with the proposed system is longer than with the traditional one. However, the proposed system could introduce more recommendations to the participants than the traditional one with no additional frustrations. It indicates that the participants can make more precise decision.

In future, we are planning to introduce characteristics into the agents. Individuality of character agents is one of the important aspects for smooth and natural interaction with human [8]. We believe it leads that conversations with the agents is more smooth and interesting.

References

1. Rist, T., Andre, E., Baldes, S.: A flexible Platform for Building Applications with Life-Like Characters. In: the 8th International Conference on Intelligent User Interfaces (IUI 2003), pp. 158–165. ACM, New York (2003)
2. Shimazu, H.: ExpertClerk: Navigating Shoppers’ Buying Process with the Combination of Asking and Proposing. In: 7th International Joint Conference on Artificial Intelligence, pp. 1443–1448 (2001)

3. André, E., Rist, T.: Adding Life-Like Synthetic Characters to the Web. In: Klusch, M., Kerschberg, L. (eds.) CIA 2000. LNCS (LNAI), vol. 1860, pp. 1–13. Springer, Heidelberg (2000)
4. Sakamoto, T., Kitamura, Y., Tatsumi, S.: A Competitive Information Recommendation System and Its Rational Recommendation Method. *Systems and Computers in Japan* 38(9), 74–84 (2007)
5. Jameson, A.: More than the Sum of Its Members: Challenges for Group Recommendation Systems. In: The Working Conference on Advanced Visual Interfaces (AVI 2004), pp. 48–54. ACM, New York (2004)
6. Mutlu, B., Shiwa, T., Kanda, T., Ishiguro, H., Hagira, N.: Footing in Human-Robot Conversations: How Robots might Shape Participant Roles Using Gaze Cues. In: the Fourth ACM/IEEE International Conference on Human-Robot Interaction (HRI 2009), pp. 61–68. ACM, New York (2009)
7. Shinohara, K., Naya, F., Yamato, J., Kogure, K.: Differences in effect of robot and screen agent recommendations on human decision-making. *International Journal of Human-Computer Studies* 62, 267–279 (2005)
8. Isard, A., Brockmann, C., Oberlander, J.: Individuality and Alignment in Generated Dialogues. In: The Fourth International Natural Language Generation Conference, pp. 25–32. Association for Computational Linguistic, Sydney (2006)

A Review of Personality in Voice-Based Man Machine Interaction

Florian Metze¹, Alan Black¹, and Tim Polzehl²

¹ LTI, Carnegie Mellon University; Pittsburgh, PA; USA

² Q&U Lab, Technische Universität Berlin; Berlin; Germany

fmetze@cs.cmu.edu

Abstract. In this paper, we will discuss state-of-the-art techniques for personality-aware user interfaces, and summarize recent work in automatically recognizing and synthesizing speech with “personality”. We present an overview of personality “metrics”, and show how they can be applied to the perception of voices, not only the description of personally known individuals. We present use cases for personality-aware speech input and/ or output, and discuss approaches at defining “personality” in this context. We take a middle-of-the-road approach, i.e. we will not try to uncover all fundamental aspects of personality in speech, but we’ll also not aim for ad-hoc solutions that serve a single purpose, for example to create a positive attitude in a user, but do not generate transferable knowledge for other interfaces.

Keywords: voice user interface, paralinguistic information, speech processing.

1 Introduction

Every speech act transmits not only a linguistic message (“text”), but it also encodes additional information in “how” things are being said. This is true for human-human communication, as well as for man-machine exchanges. In this paper, we will discuss the role of “personality” in voice-based user interfaces, and the state-of-the-art for implementing systems that can recognize and synthesize personality features. We will discuss ways to analyze speech data recorded from humans, advanced speech synthesis methods, and ways to encode personality information in the original text message, which is being transmitted. We believe that future research on voice-enabled human computer interfaces must go beyond the analysis of “what” is being said, and should include aspects of “how” it is being said. This capability is needed in order to adapt to an unknown user, or a known user’s changing state of mind. The system must then send consistent messages across all channels used, i.e. it would express the same message by using different words, and by modifying the voice characteristics used.

Personality is certainly the richest means for characterizing, and even classifying people [16], and people assign it rapidly and automatically [18]. This instinct allows us humans to quickly construct a model of a person we meet, and predict a wide range of attitudes, behavior, and other properties, which we expect to encounter during interaction. Personality can for example be used to differentiate the introverted from

the extroverted, the shy from the exuberant, the egoist from the altruist, or the conservative from the adventurous. We will assume that extroverted persons talk more than they listen, and use strong language, while introverted persons listen more, and use qualifiers such as “maybe”, or “perhaps”. The concept of personality gives us cues on what to expect from others, and how to behave ourselves. Descriptions such as “extroverted” or “introverted” serve as a shorthand descriptor for a bundle of traits, which we attribute to persons. Interestingly however, self-reporting of personality traits often leads to different results than attribution by others.

2 Personality in Voice-Based Man Machine Interaction

In an advanced voice user interface, the computer should be aware of the human’s personality and tailor its response accordingly. Similarly, the user’s behaviour will be influenced by his perception of the system’s personality, conveyed by what the system says, and how it is being said. In the “Computers as Social Actors” (CASA) paradigm, Nass and Brave [16] postulate that humans communicate with machines just as they would with another human. Generally, when people encounter someone who seems to have a personality like their own, they tend to have positive feelings toward that person [24]. They conclude that designers of user interfaces should therefore seek to manipulate the speech characteristics of any technology that can produce speech, and thereby give it a personality. If one wants to be able to adapt to unknown users, possibly in real-time, the human’s speech should also be analysed for personality traits, as should other input channels. In an automated voice user interface, the assessment of a user’s personality must be done within seconds, and on the basis of the speaker’s voice only. Methods established and verified in psychology, like the use of long questionnaires [7], are therefore inapplicable, or at least cumbersome.

In a more general setting, personality is also a property of embodied conversational agents (ECAs) [4]. Cassell et al. show that perceived personality of the agent is a major factor in the perception of such user interfaces [2]. Catrambone et al. list the personality of the user as one of the factors to be included in an evaluation of ECAs [5], arguing that an understanding of the mechanisms involved will eventually allow the design of appropriate personalities. [6] presents on-going work which uses a Wizard-of-Oz paradigm, because personality can not currently be analysed and synthesised satisfactorily using fully automatic means, as would be required.

Given the above it is clear that personality must be modelled properly in the audio channel of any speech-enabled multi-modal user interface.

While we have used the term “personality” many times already in this paper, we have not formally defined it yet. Following Ryckman [25], personality can be defined as *“a dynamic and organized set of characteristics possessed by a person that uniquely influences his or her cognitions, motivations, and behaviours in various situations.”* In our own work, we follow the trait theory of personality [12], and see personality as a defined set of habitual patterns of behaviour, thoughts, and emotions. We can then apply an assessment scheme using the “Big Five” NEO-FFI [7] personality traits. We chose this scheme, because the traits are seen as empirical observations, not a fundamental theory, which aims to fully explain personality. [9] gives an overview of different schools of describing the rich concept of personality.

The NEO-FFI describes personality traits along five ordinal dimensions, which are called “scales”, namely *Neuroticism (N)*, *Extroversion (E)*, *Openness (O)*, *Agreeableness (A)*, and *Conscientiousness (C)*. Human raters generate another person’s profile by giving answers to 60 propositions from the NEO-FFI questionnaire (called “items”) using a 5-point Likert scale ranging from “strongly disagree” to “strongly agree”. A self-report form is also available, but was not used for the experiments described here. The 60 items are then aggregated into numeric values for the 5 scales using the NEO-FFI coding scheme. The questionnaires and the resulting scales and factors have been validated with high consistency, including translations, cross-cultural experiments and retests, confirming the reliability of this approach for a large number of conditions. The German NEO-FFI, which was used in our experiments described below, has been validated with more than 12.000 test persons.

In the context of a voice-based communication, the scales correspond to vocal manifestations of *perceived* personality traits, unless the participants have other cues on which to base their judgement, for example previous, external knowledge, or the transmission of a message in another, conflicting personality (see Section 0). Attribution happens on basis of auditory impressions, and it is questionable how this compares to the conventional assessment methods, where raters know the person to be rated. In studies, Nass et al. find low, but significant correlation for synthetic speech [17]. This shows that personality impressions can be generated by the choice of a voice, and that they can influence the perception of other information presented at the same time. Our results below confirm this conclusion.

3 Recognizing Personality in Speech

Apple et al. [1] found that pitch and speaking rate influence the perception of speakers’ voice with regard to factors such as truthfulness, empathy, “potency”, amongst others. They also observed interplay between the message (the text which was spoken) and the effect of a manipulation of the above factors towards the attribution. Scherer & Scherer [26] analysed prosodic features such as pitch and intensity, and observes that extroverted speakers speak louder, and with fewer hesitations. They suggest that extroversion is the only factor that can be reliably estimated from speech. Mairesse [15] also finds that prosodic and acoustic features are important cues for recognizing extroversion, and that extroversion can be modelled best, followed by emotional stability (neuroticism) and openness to experience. Finally, Nass and Lee established that humans could infer personality impressions even from automatically synthesized speech [17]. They find that humans are attracted more to a voice that is similar to their own, and that it is possible to generate extroverted and introverted synthetic voices, which people will recognize as such. Our recent, more systematic work roots in previous experiments on emotion recognition [23], which also showed the benefit of relying on multiple information sources, acoustic and linguistic cues in this case. In [22], we present results of an automatic assessment of all five NEO-FFI traits.

3.1 Database and Human Perception of Personality in Speech

We recorded the “natural” voice of a professional speaker, who had previously recorded voice prompts in speech dialog systems, and was used to working with voice

coaches. We then presented him the original descriptions of the 5 NEO-FFI personality traits from the NEO-FFI manual. He prepared 10 voice personalities, representing persons with either high or low values on each of the five scales. We therefore have 11 different recording conditions: 2 extremes on each of the five scales, plus “normal”.

The spoken text is designed to resemble a neutral, complete phrase, typical for a hotline, which comprises a welcome, information on a voucher redeemer service, and a short goodbye. Each recording lasts about 20s. We recorded at least 20 takes of each of the conditions, more than an hour of speech in total. All speech samples were annotated for “artificiality” by two experienced labellers, and we retained for our human rating experiments the three least artificial takes for each condition.

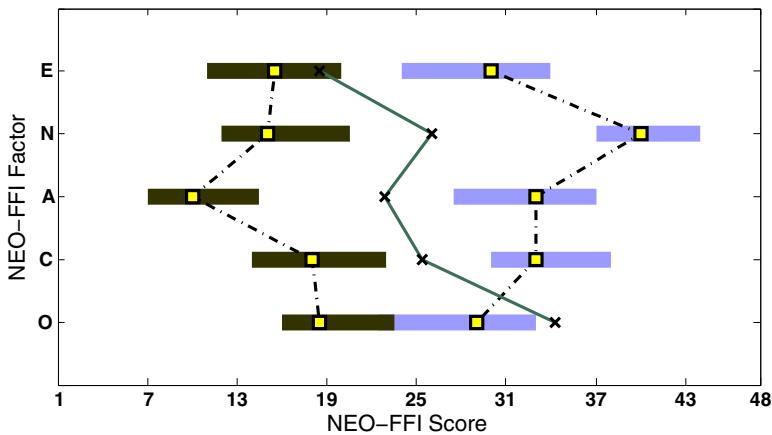


Fig. 1. NEO-FFI ratings for speech: brown bars (left) represent inter-quartile ranges of ratings from variation towards low values, light blue bars (right) towards high values. Vertical lines connect the medians, i.e. solid line for “normal personality”, dashed for acted variations.

87 raters rated 8 takes from different conditions on average. 20 different raters rated every take. Raters could listen to the takes through high-quality headsets up to 5 times, while completing a NEO-FFI questionnaire about their impression of this take’s speaker. Overall, over 600 questionnaires for all 5 scales were generated.

Fig. 1 shows the distribution of the raters’ assessments for both the acted and the natural speech samples for the 5 factors. Each data point represents 60 ratings from 3 different takes. Overall, raters were able to label the acted personalities quite well, as nearly all the conditions were perceived as intended by the actor. In our recordings, the speaker successfully varied the values of the factors *N*, *C*, and *A*, while *E* and *O* seem more difficult. While the attempt to lower the perceived extroversion in speech had only little effect, the attempt to raise the impression of openness in fact lowered the perceived score. This could be due to the “natural” value for this speaker being quite extreme already for *E* and *O*, an inability of our particular speaker to act these traits, or a general difficulty in perceiving and assessing these modifications from speech, or our speech sample. Further experiments will be needed to answer these

questions. These findings coincide with [26] and [15], and show that impressions of extroversion and neuroticism can be distinguished using speech. Furthermore, we show for our speaker that all 5 traits can be varied and recognized. We also observe low differences between the distribution variances of ratings on different conditions.

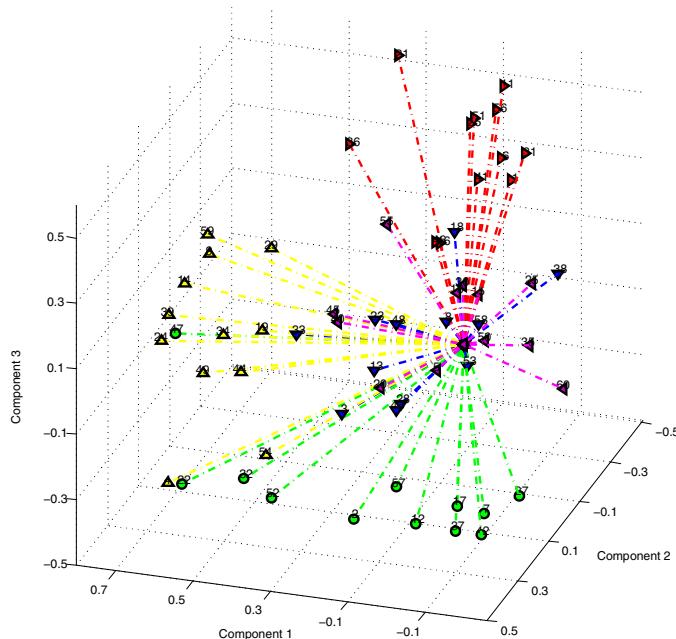


Fig. 2. NEO-FFI loadings of the factor scores, shown in a three-dimensional projection on the user-test assessment space. The original loadings, established using the questionnaire for known persons (marked by different colors), are well reproduced in the assessments of acted speech, as the colored “bundles” generally point in the same directions.

In order to compare the factors in the traditional NEO-FFI coding scheme with the underlying structure of the ratings on our recorded speech data, we conducted an exploratory factor analysis [8], hypothesizing the presence of 5 latent factors in the user ratings. Figure 2 shows the 3 most dominant (latent) factors, in the given NEO-FFI coding space. Lines correspond to directions of item loadings in our data. Colours correspond to item loading membership in the NEO-FFI coding scheme. The coding found in our data correlates quite well with the NEO-FFI coding scheme, as most factors with the same colour (original loadings) point in a similar direction (new factor loadings). We also observe a low number of cross-loadings between factors, and low to moderate commonality for almost all items [22].

In sum, this experiment shows that the NEO-FFI scheme can not only be used for assessment of known persons’ personalities, but also to create profiles of perceived personality, from listening to short samples of speech.

3.2 Automatic Recognition of Personality from Speech

Our automatic system computes and classifies prosodic and acoustic speech properties. The feature set leverages previous work on emotion recognition [23]. We extract audio descriptors such as 16 MFCC coefficients, 5 formant frequencies, intensity, pitch, perceptual loudness, zero-crossing rate, harmonics-to-noise-ratio, centre of spectral mass gravity (centroid), the 95% roll-off point of spectral energy and the spectral flux, etc., using a 10ms frame shift. From these descriptors, we derive statistics at the utterance level, separate for voiced and unvoiced regions, on speech parts only. These statistics include means, moments of first to fourth order, extrema, skewness, kurtosis, and ranges from the temporal contours over one utterance. To model temporal behaviour, we append first and second order finite differences. In total, 1450 features are being computed. Before classification or regression using a Support Vector Machine (SVM) with linear kernel functions [30], the most salient features are being selected by ranking them according to Information Gain Ratio (IGR) using an Sequential Floating Forward Selection (SFFS) wrapper, and only retaining the top N features in a given set.

Using 10-fold cross-validation, we obtain an accuracy of approximately 60% on our balanced ten-class classification task, consisting of the high and low targets for the 5 personality traits, which is six times the chance level. Because humans have only been asked to fill in NEO-FFI questionnaires, and did not perform a classification task, a human baseline cannot be computed. Interestingly, high and low neuroticism and conscientiousness, as well as high extroversion can be recognized better than the other classes, with class-specific F -measures between 0.70 and 0.89, while the other classes perform between 0.32 and 0.54. Best results were achieved when using about 40 features, although very little change occurs as soon as at least 20 features are being retained. High extroversion (E) can be classified well, which is in line with observations by [26, 15]. Most problematic are the O and A factors. Different from separability by humans (see Figure 1), automatic classification gives poor results for A . O seems to be hard for both human and automatic classification.

Analysing the most salient feature types, we observe a predominance of MFCC-based features. Most important are the statistics derived from the unvoiced speech parts. Also features from intensity and duration of segments, as well as pitch derivatives are of high importance, e.g. the maximum intensity from unvoiced speech parts or the distribution and percentage of voiced segments overall. Features capturing dynamics of unvoiced speech parts are generally sensitive to strength of fricatives and plosives. Features capturing pitch variation and derivatives are generally sensitive to intonation movements. Along with cepstral features, which also partly capture spectral sharpness and tilt, the importance of these features seem to be in line with results from auditive analyses presented in earlier work.

In many applications, however, we may not be interested in automatic classification of personality, or in a personality assessment by a machine, but in the reproduction of a human rating by a machine.

We therefore conducted a regression experiment, in which we use the (numeric) ratings of the labellers as ground truth. In this experiment, we use all available ratings for the speech recordings, and SVM regression. Correlation analysis shows how different the predictions by humans and machines are, for the various factors. As in

the classification experiments, there is almost no change when using more than 40 features. Analysing the top ranked features, we see that for factors O and C , predominantly MFCC features are being used. For the other factors, the picture seems much more diverse. For factors E and A , features that capture dynamics of pitch are given high ranks, e.g. standard deviation, slopes, ranges, derivatives. For N , loudness and intensity features are prevalent, using statistics describing the distribution, e.g. skewness or kurtosis. Interpreting our results, degrees of extroversion and agreeableness seem to be conveyed much more by tonal expression than degrees of other factors. In addition, intensity and loudness levels can be exploited to gain indications of vocal impression of neuroticism. Further research will focus on a detailed interpretation of these findings. Generally, our findings are again in line with previous work on signal-based analysis [26, 15].

Comparing results from classification and regression analysis, we observe that predicting factors values and classifying for binary classes can be applied with good results for factors neuroticism (N) and extroversion (E). While classifying into high and low variations along the conscientiousness (C) dimension also yields reasonable classification scores, our models poorly predict the value humans would assign to that factor. Relatively poor results are achieved for openness (O) and agreeableness (A).

4 Synthesizing Voices with Personality

The messenger influences our perception of the message [16]. For isolated experiments in lab settings, it may be enough to manually control volume, pitch, pitch range, and speech range in a desired way, but by now a significant body of work in “expressive” synthesis exists, which allows to systematically modify voice properties associated with emotions, or personality. In practice, volume is very hard to control in real-world settings, for example over a telephone line.

A number of groups have been successful at synthesizing different emotional states, which is similar to synthesizing speech with personality. This is most often done by using acted data of different states, and using models trained from that data to impose prosody (intonation, duration and phrasing) on synthetic output. It is possible to classify such synthesis attempts into two forms following the current two major techniques in speech synthesis.

Using unit selection [13] technology, where appropriate sub-word units are selected from large natural speech database, requires that the database itself contains examples of the desired emotional state. [10] achieved this by deliberately recording different versions of their database with the range of desired emotional states. This did have some success, but it was of course limited to the types of data recorded, and to some extent the domain of the recorded data. [28] however take a more indirect approach. For a spoken dialog system, instead of recording prompts in isolation with explicitly stated emotional variation, they recorded the prompts within an actual simulated spoken dialog, where the voice talent plays the machine end. In their work they find the voice talent modified their prosody naturally given the dialog context. Also, by adding a prosody feature based on the classification of different dialog states, they could improve synthesis. But again this technique is very targeted toward the particular application and dialog context that the synthesizer would be used in.

The second synthesis technique is Statistical Parametric Speech Synthesis [31], where speech data is modelled in a generative fashion, which as a first approximation can be viewed as averaging, in contrast to clustering instances of data in the unit selection case. Statistical Parametric Speech Synthesis is typically using smaller amounts of data, thus can produce a wider range of output than a unit selection system could on the same data. [3] describes using statistical models of F_0 power and duration to model different emotional states. But even though statistical parametric speech synthesis allows for more control of the modelling, it is still hard to get the distinctions significant enough to allow the listener to perceive the intended state.

[29] propose an evaluation strategy for expressive speech, but admit that it is hard to evaluate all the subtle variations. It is possible to make general remarks about prosody use in stylistic speech. Angry and happy speech typical has higher F_0 s and larger dynamic range. Sad speech typically has lower F_0 and small dynamic range and longer durations. However more subtle differences are harder to explicitly describe, and synthesis from them equal harder to do well. Though in the similar problem of voice conversion, which can be used for both conversion to new voices but also conversion to new styles from the same speaker, we have found it useful to train Speaker ID classifiers and use them to evaluation metrics for our voice conversion [14]. Assuming that the speaker ID (or personality ID) classifiers are trained on human speech, if synthesized examples can be classified in the intended way then we have a good evaluation technique (and optimization measure) for building synthesis models. However it has been noted that even when objective measures are satisfied, that does not mean human evaluator necessary agree.

5 Consistency

Outside of speaker identification and verification, there will hardly ever be a voice-based user interface, in which the voice itself is the only information being required and exchanged. It will therefore be important to convey the same notion of personality in these modalities during synthesis.

Examining distinctive linguistic and lexical choices in 2007, Gill [11] investigates the relationship between the personality of an author of short emails and blog texts, generated by self-assessment, and their language. He observes weak correlations, but concludes that personality is represented in text using more complicated features. Oberlander [19] examines the relation between part-of-speech (POS) distributions in email texts and two distinct personality traits, neuroticism and extroversion, of their authors. He concludes that part-of-speech information can be characteristic.

During generation of text using a binary extroversion/ introversion distinction, Nass et al. show that relatively simple rules, for example replacing weak adjectives and quantifiers (“quite rich”) by strong language (“absolutely sensational”), will change the perception of the text, particularly when paired with a corresponding synthetic voice [17]. In fact, the voice properties seem to have a stronger effect than the text properties, in particular when matched to the reader/ listener.

6 What to Do? What Next?

While there won't be a simple recipe to go by for quite some time yet, we believe that speech technology will soon be able to automatically recognize and generate more complex personality structures than just a binary "introverted vs. extroverted" distinction. More detailed, and reliable, automatic analysis of voice signals, be they synthetic, natural, or concatenated, will be necessary for this type of technology to leave the lab. While it is reasonable to expect to be able to manipulate the "volume" of a voice prompt in a lab setting, it can be quite difficult to do the same thing over the telephone, where a variety of network transmission channels and handsets are waiting, with no control on the part of the service provider.

While the use of personality in user interfaces may well be guided by relatively simple rules ("similar attract") for quite some time, a company might for example want to make sure that voices that have been casted for brand image are perceived the same on high-quality head-sets and on low-grade equipment with different transmission characteristics (imagine long-distance lines or VoIP connections). Once the correlation between human ratings (perception) and factors, which can be varied during synthesis (generation) are known, simple rules can be replaced by algorithms, with parameters that can be fitted to specific situations, ensuring appropriate service.

Of course, the speech features presented here will only be a fraction of the information, which can be extracted from a multi-modal interface; we have neglected text, timing, video/ graphical, or haptic information, if available. All of these have to be analysed together, and synthesised consistently. Still, meta-data extraction from speech is currently a very active field of research with challenge-style evaluations [27], as is the "social signal processing" community [20]. We hope that the significant effort directed towards automatic analysis of meta-data in speech will soon uncover relevant factors and features in more detail, which can then be used to automatically generate appropriate speech. We believe that the ability to model personality in speech recognition and synthesis will be a big and necessary step towards natural-like man machine interaction, as promised by the vision of "affective computing" [21].

References

- [1] Apple, W., Streeter, L.A., Krauss, R.M.: Effects of pitch and speech rate on personal attributions. *Journal of Personality and Social Psychology* 37(5), 715–727 (1979)
- [2] Bickmore, T., Cassell, J.: Social Dialogue with Embodied Conversational Agents. In: *Natural, Intelligent and Effective Interaction with Multimodal Dialogue Systems*. Kluwer Academic, New York (2004)
- [3] Bulut, M., Lee, S., Narayanan, S.: A statistical approach for modeling prosody features using postags for emotional speech synthesis. In: Proc. ICASSP, Honolulu, HI (2007)
- [4] Cassell, J., Sullivan, J., Prevost, S., Churchill, E.F. (eds.): *Embodied Conversational Agents*. MIT Press, Cambridge (2000)
- [5] Catrambone, R., Stasko, J., Xiao, J.: Anthropomorphic agents as a user interface paradigm: Experimental findings and a framework for research. In: Proc. 24th Annual Conference of the Cognitive Science Society, Fairfax, USA (August 2002)
- [6] Chen, Y., Naveed, A., Porzel, R.: Behavior and preference in minimal personality: A study on embodied conversational agents. In: Proc. ICMI-MLMI. ACM Press, New York (2010)

- [7] Costa, P.T., McCrae, R.R.: Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) manual. Psychological Assessment Resources (1992)
- [8] Costello, A.B., Osborne, J.W.: Best practices in exploratory factor analysis. *Practical Assessment, Research & Evaluation* 10(7) (July 2005)
- [9] Drapela, V.J.: A Review of Personality Theories, 2nd edn. Charles C. Thomas Publ. (1995)
- [10] Eide, E., Bakis, R., Hamza, W., Pitrelli, J.: Multilayered extensions to the speech synthesis markup language for describing expressiveness. In: Proc. Eurospeech, Geneva, Switzerland (2003)
- [11] Gill, A.J., French, R.M.: Level of Representation and Semantic Distance: Rating Author Personality from Texts. In: Proc. Euro Cogsci, Delphi, Greece (2007)
- [12] Goldberg, L.R.: The structure of phenotypic personality traits. *American Psychologist* 48, 26–34 (1993)
- [13] Hunt, A., Black, A.: Unit selection in a concatenative speech synthesis system using a large speech database. In: Proc. ICASSP, Atlanta, Georgia, vol. 1 (1996)
- [14] Jin, Q., Toth, A., Black, A., Schultz, T.: Is voice transformation a threat to speaker identification? In: Proc. ICASSP, Las Vegas, USA, NV (2008)
- [15] Mairesse, F., Walker, M.A., Mehl, M.R., Moore, R.K.: Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. *Journal of Artificial Intelligence Research (JAIR)* 30, 457–500 (2007)
- [16] Nass, C., Brave, S.: *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship*. MIT Press, Cambridge (2005)
- [17] Nass, C., Lee, K.M.: Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied* 7, 171–181 (2001)
- [18] Nass, C., Moon, Y., Fogg, B., Reeves, B., Dryer, D.C.: Can computer personalities be human personalities? *International J. of Human-Computer Studies* 43(2), 223–239 (1995)
- [19] Oberlander, J., Gill, A.J.: Individual Differences and Implicit Language: Personality, Parts-of-Speech and Pervasiveness. In: Proc. Cogsci, Chicago, IL, USA (2004)
- [20] Pentland, A.: Social signal processing. *IEEE Signal Proc. Magazine* 24(4), 108–111 (2007)
- [21] Picard, R.W.: *Affective Computing* (1995)
- [22] Polzehl, T., Möller, S., Metze, F.: Automatically assessing acoustic manifestations of personality in speech. In: Proc. SLT Workshop. IEEE, Berkeley (2010)
- [23] Polzehl, T., Schmitt, A., Metze, F., Wagner, M.: Anger recognition in speech using acoustic and linguistic cues. *Speech Communication, Special Issue on Sensing Emotion and Affect - Facing Realism in Speech Processing* (2011)
- [24] Reeves, B., Nass, C.: *The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places*. Cambridge University Press, Cambridge (1996)
- [25] Ryckman, R.M.: *Theories of Personality*. Thomson/Wadsworth, Belmont CA (2004)
- [26] Scherer, K.R., Scherer, U.: Speech Behavior and Personality. *Speech Evaluation in Psychiatry*, 115–135 (1981)
- [27] Schuller, B., Steidl, S., Batliner, A.: The INTERSPEECH 2009 emotion challenge. In: Proc. INTERSPEECH, ISCA, Brighton, UK (September 2009)
- [28] Syrdal, A., Conkie, A., Kim, Y., Beutnagel, M.: Speech acts and dialog TTS. In: Proc. SSW 7, Keihanna, Japan (2010)
- [29] Türk, O., Schröder, M.: Evaluation of expressive speech synthesis with voice conversion and copy re-synthesis techniques. *IEEE Trans. on ASLP* 18(5), 965–973 (2010)
- [30] Witten, I.H., Frank, E., Trigg, L., Hall, M., Holmes, G., Cunningham, S.J.: *Weka: Practical machine learning tools and techniques with java implementations* (1999)
- [31] Zen, H., Tokuda, K., Black, A.: Statistical parametric speech synthesis. *Speech Communication* 51(11), 1059–1064 (2009)

Can Indicating Translation Accuracy Encourage People to Rectify Inaccurate Translations?

Mai Miyabe¹ and Takashi Yoshino²

¹ Graduate School of Systems Engineering, Wakayama University,
930 Sakaedani, Wakayama, Japan
miyabe@yoslab.net

² Faculty of Systems Engineering, Wakayama University,
930 Sakaedani, Wakayama, Japan
yoshino@sys.wakayama-u.ac.jp

Abstract. The accuracy of machine translation affects how well people understand each other when communicating. Translation repair can improve the accuracy of translated sentences. Translation repair is typically only used when a user thinks that his/her message is inaccurate. As a result, translation accuracy suffers, because people's judgment in this regard is not always accurate. In order to solve this problem, we propose a method that provides users with an indication of the translation accuracy of their message. In this method, we measure the accuracy of translated sentences using an automatic evaluation method, providing users with three indicators: a percentage, a five-point scale, and a three-point scale. We verified how well these indicators reduce inaccurate judgments, and concluded the following: (1) the indicators did not significantly affect the inaccurate judgments of users; (2) the indication using a five-point scale obtained the highest evaluation, and that using a percentage obtained the second highest evaluation. However, in this experiment, the values we obtained from automatically evaluating translations were not always accurate. We think that incorrect automatic-evaluated values may have led to some inaccurate judgments. If we improve the accuracy of an automatic evaluation method, we believe that the indicators of translation accuracy can reduce inaccurate judgments. In addition, the percentage indicator can compensate for the shortcomings of the five-point scale. In other words, we believe that users may judge translation accuracy more easily by using a combination of these indicators.

Keywords: multilingual communication, machine translation, back translation.

1 Introduction

The Internet has increased the opportunity for multilingual communication. However, communicating in non-native language is complicated, because the language barrier hampers mutual understanding [1, 2]. Machine translation is used in order to overcome the language barrier when communicating using non-native language [3].

Despite recent advances in machine translation technology, obtaining highly accurate translations is still very difficult. The probability of inaccurate machine translations increases as messages become longer. As we have pointed out, inaccurate

translations impede mutual understanding. Therefore, for smooth communication, users need to create messages with very few translation errors.

Translation repair plays an important role in multilingual communication when using machine translation, as it can be used to create messages with very few translation errors [4]. Translation repair is typically only performed when a user assumes that his/her message is inaccurate. As a result, translation accuracy suffers, because people's judgment in this regard is not always accurate, and many inaccurate messages are not repaired effectively [5]. Therefore, it is necessary to develop a method that helps to reduce inaccurate judgments on the part of users.

There are differences among users' judgments of translation accuracy. Inaccurate judgments may occur because of irresponsible users. We believe that people can more correctly judge the accuracy of a translation if they are provided with an indicator of the accuracy of that translation. In this study, we propose a method that indicates the accuracy of a translation in order to reduce users' inaccurate judgment. In particular, the method uses an automatic method to evaluate the accuracy of the translation, and then indicates this value to the user. This paper verifies the effect of indicating the translation accuracy using the proposed method.

2 Translation Accuracy Indication of Back-Translated Sentences

In order to indicate translation accuracy, we have to measure the accuracy of the translated sentences. In this study, we use an automatic evaluation method of translation accuracy, called Bilingual Evaluation Understudy (BLEU). This method is one of a number of automatic evaluation methods for translation accuracy. In the area of natural language processing, various studies use BLEU as an automatic evaluation method.

Many researchers have proposed different automatic evaluation methods of translation accuracy, including BLEU [6], NIST [7], and so on. These methods calculate the accuracy of a translation by comparing a translated sentence to human reference translations. Uchimoto et al. proposed a method that calculates translation accuracy by comparing a back-translated sentence with its input sentence [8]. We assume that people use back translation for creating multilingual messages. Therefore, we use the latter method here.

2.1 Accuracy Indicators

In this study, we propose the following three methods for indicating translation accuracy.

Method (A): Translation accuracy is expressed in percentage terms.

Method (B): Translation accuracy is expressed using the following five-point scale¹: "It is correctly translated," "It is translated correctly, but is not fluent," "It is not entirely comprehensible," "It is partially translated, but does not express the meaning of the input sentence," and "It is incorrectly translated."

Method (C): Translation accuracy is expressed using the following three-point scale: Correct, Neutral, and Wrong.

¹ We defined the five-point scale on the basis of the adequacy evaluation method developed by Walker [9].

2.2 Evaluation Method of Translation Accuracy

In this study, we use BLEU [6] to measure translation accuracy. We derive the values for each of the three indicators from the BLEU score. The BLEU score is calculated using the following equations:

$$\text{BLEU} = \text{BP} \times \exp(\log P) \quad (1)$$

$$\text{BP} = \begin{cases} 1 & (c \geq r) \\ \exp(1 - r/c) & (c < r) \end{cases} \quad (2)$$

$$P = \frac{N_s}{c} \quad (3)$$

Here,

c is the number of words in a back-translated sentence.

r is the number of words in an input sentence.

N_s is the number of matching words between an input sentence and its back-translated sentence.

A BLEU score gives a value between 0 and 1. Therefore, using method (A), the BLEU score is provided to users. For method (B), we convert the BLEU score to values on a five-point scale using the following equation.

$$\text{Value} = 2.856 \times \text{BLEU} + 1.183 \quad (4)$$

Table 1 shows the relationship between the BLEU score, and the values of the five-point scale and the three-point scale. For method (C), we convert the values on the five-point scale to values on a three-point scale using the relationship on table 1.

Table 1. The relationship between the BLEU score, the five-point scale, and the three-point scale

BLEU score	Five-point scale	Three-point scale
$1 \leq \text{Value} < 2$	“It is incorrectly translated.”	Wrong
$2 \leq \text{Value} < 3$	“It is partially translated, but does not express the meaning of the input sentence.”	
$3 \leq \text{Value} < 4$	“It is not entirely comprehensible.”	Neutral
$4 \leq \text{Value} < 5$	“It is translated correctly, but is not fluent.”	Correct
$\text{Value} = 5$	“It is correctly translated.”	

3 Experiment

We performed an experiment to verify the effect of each indicator. The subjects of the experiment were 20 university students. Their ages ranged from 18 to 24 years, with an average age of 21 years.

In this experiment, we used 40 Japanese sentences containing conversational expressions. These sentences required translation repair because their back-translated sentences were inaccurate. We decided to limit the number of letters in a sentence to between 20 and 30.

3.1 Evaluation Points

The points of evaluation were as follows:

[**Point 1**] Can indicating translation accuracy encourage people to rectify inaccurate translations?

[**Point 2**] Which method is most effective?

3.2 Experimental Condition

In order to verify the evaluation points, the experiment was performed under the following four conditions:

[**Condition 1**] Without a translation accuracy indicator

[**Condition 2**] With the translation accuracy indicator using Method (A)

[**Condition 3**] With the translation accuracy indicator using Method (B)

[**Condition 4**] With the translation accuracy indicator using Method (C)

The subjects repaired 10 translation sentences under each condition.

3.3 Experimental Tool

Figure 1 shows a screenshot of our experimental tool. When a subject inputs a sentence into the input area, its back-translated sentence is shown in the back translation area. Under experimental conditions 2, 3, and 4, the value of the calculated accuracy indicator is shown at the end of the back-translated sentence. The calculated accuracy indicator is highlighted in red.

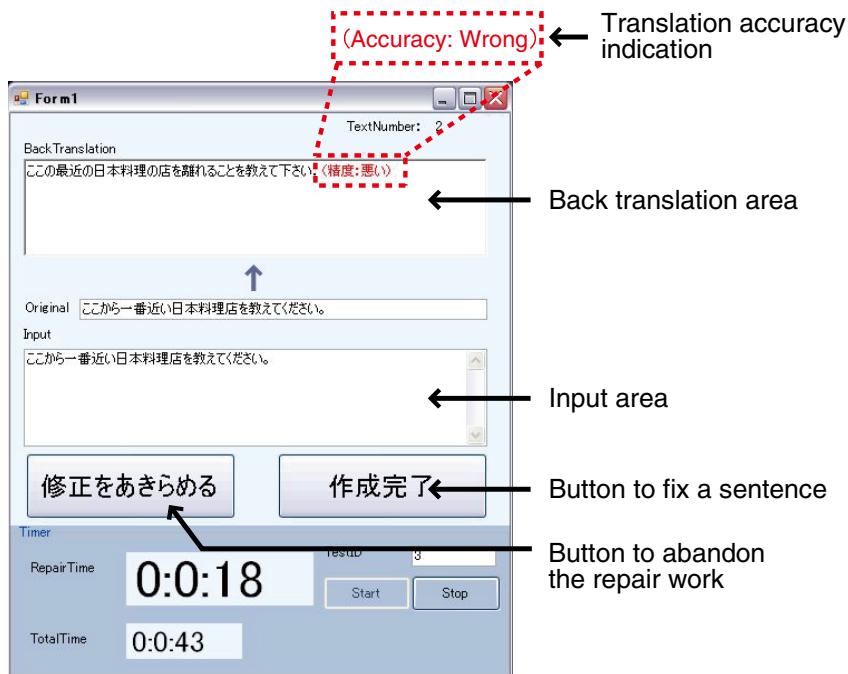
This experimental tool used a J-Server with Language Grid [10] as the machine translation system. In this experiment, the source language for the translation was Japanese, and the target language was Chinese. In order to calculate the BLEU score, this tool carries out the Japanese morphological analysis using the morphological analyzer MeCab [11].

3.4 Experimental Procedure

The experimental procedure was as follows:

- (1) Repair an input sentence to ensure that the meaning of the back-translated sentence conforms to that of the input sentence.
- (2) Fix the back-translated sentence when a subject concludes that its meaning is the same as that of the input sentence.
- (3) Repeat steps (1) and (2) 10 times.
- (4) Repeat steps (1) to (3) for the other experimental conditions.

When the subjects were unable to repair a sentence after 5 min, we allowed them to abandon the translation repair.

**Fig. 1.** Screenshot of experimental tool**Table 2.** Evaluated accuracy in each experimental condition

	Unrepaired sentences	Repaired sentences			
		Condition 1 (Without indication)	Condition 2 (Method (A))	Condition 3 (Method (B))	Condition 4 (Method (C))
Average	1.6	3.2	3.2	3.2	3.3
S.D.	0.5	0.6	0.5	0.6	0.7
Significance probability	<0.001				

Table 3. Number of the sentences abandoned in the experiment

	Condition 1 (Without indication)	Condition 2 (Method (A))	Condition 3 (Method (B))	Condition 4 (Method (C))
Number of sentences abandoned	41 sentences	34 sentences	40 sentences	51 sentences
Abandonment rate	20.5 %	17.0 %	20.0 %	25.5 %

In each condition, the total number of repaired sentences was 200.

4 Results

4.1 Accuracy of Repaired Back-Translated Sentences

We evaluated the accuracy of both the unrepaired back-translated sentences and the repaired back-translated sentences by using the adequacy evaluation method developed by Walker [9]. In this method, a five-point scale is used to evaluate the translation accuracy. The evaluation method asks the question: "How much of the meaning expressed in the input message is also expressed in the back-translated message?" The method then grades the accuracy value on the following scale: 5: All, 4: Most, 3: Much, 2: Little, and 1: None. In this evaluation, three evaluators evaluated the accuracy of back-translated sentences.

Table 2 shows the evaluated accuracy in each experimental condition. As can be seen from the results of the evaluation, there was a significant difference among unrepaired sentences and repaired sentences under the four conditions. In addition, we see from the results of the multiple comparisons that there was no significant difference between unrepaired sentences and repaired sentences under each condition.

4.2 Number of Sentences Abandoned

Table 3 shows the number of the sentences abandoned in the experiment. The purpose of this study is to verify how well the indicators reduce the number of inaccurate judgments. The subjects abandoned the repair work when they had not improved the accuracy of a sentence enough after 5 minutes. Therefore, we consider that the abandoned sentences occurred as a result of the subjects' accurate judgments.

Table 4. Number of the inaccurate judgments

	Condition 1 (Without indication)	Condition 2 (Method (A))	Condition 3 (Method (B))	Condition 4 (Method (C))
Number of inaccurate judgments	51 sentences	58 sentences	49 sentences	40 sentences
Inaccurate judgment rate	25.5 %	29.0 %	24.5 %	20.0 %

In each condition, the total number of repaired sentences was 200.

Table 5. Results of questionnaire

Questions	Average (S.D.)
(1) I checked the translation accuracy when the accuracy was indicated.	4.2 (1.1)
(2) I think that the method (A) was useful for judgments of translation accuracy.	3.9 (1.2)
(3) I think that the method (B) was useful for judgments of translation accuracy.	3.9 (0.9)
(4) I think that the method (C) was useful for judgments of translation accuracy.	2.9 (0.9)

We used a five-point Likert scale for the evaluation: 1: Strongly disagree, 2: Disagree, 3: Neutral, 4: Agree, and 5: Strongly agree.

Table 6. Frequency distribution of ranking of indicators by subjects

	Method (A)	Method (B)	Method (C)
Rank 1	8	11	1
Rank 2	9	7	4
Rank 3	3	2	15
Σar	35	31	54

Σar is the sum of the values calculated by multiplying the frequency of appearance by the rank value.

4.3 Number of Inaccurate Judgments

In this experiment, we evaluated the accuracy of the repaired back-translated sentences using a five-point scale. If the accuracy evaluation is greater than or equal to 3, the back-translated sentence is passed on. If the value is less than 3, the back-translated sentence is not passed on. Therefore, if the accuracy of a repaired sentence is less than 3, we regard it as an inaccurate judgment. We counted the number of inaccurate judgments of 200 repaired sentences.

Table 4 shows the number of the inaccurate judgments. From table 4, we see that Condition 4 had the least number of the inaccurate judgments. Similarly, Condition 2 had the most number of inaccurate judgments.

4.4 Results of Questionnaire

Table 5 shows the result of the questionnaire. We used a five-point Likert scale for the evaluation: 1: strongly disagree, 2: disagree, 3: neutral, 4: agree, and 5: strongly agree.

Moreover, in the questionnaire, we asked subjects to rank the three methods for indicating translation accuracy. Table 6 shows the results of the ranking. The smaller the value of Σar in table 6, the better is the ranking. From the results of the ranking, we see that method (B) has the smallest Σar , and method (C) has the largest Σar . Therefore, according to the subjects, method (B) was evaluated as the best method for indicating translation accuracy.

4.5 The Accuracy of the Automatic Evaluation Method

In this experiment, the experimental tool indicated the accuracy of back-translated sentences calculated using the equations described in section 2.2. However, the calculated accuracy is not necessarily accurate. Therefore, we verified the correlation between a subjective evaluation and the BLEU score.

The correlation coefficient between the subjective evaluation and the BLEU score was 0.335, and the significance probability was less than 0.001. Although there was a positive correlation between the subjective evaluation and the BLEU score, it was not very high, and so there may have been a mismatch between them.

Table 7. Number of the inaccurate judgments with an incorrect indication of translation accuracy

	Condition 3 (Method (B))	Condition 4 (Method (C))
Number of inaccurate judgments	23 sentences	18 sentences

5 Discussion

5.1 Reducing the Effect of Inaccurate Judgments

In this section, we discuss evaluation point 1: “Can indicating translation accuracy encourage people to rectify inaccurate translations?”

In section 4.3, we showed that Condition 4 had the least number of the inaccurate judgments and that Condition 2 had the most inaccurate judgments. However, the difference between the highest number and the lowest number was 18. Therefore, there was no large difference between each of the conditions.

In the free description section of the questionnaire, subjects commented that “with translation accuracy indications using methods (B) and (C), I finished repair work when the experimental tool indicated ‘It is not entirely comprehensible’ or ‘Neutral’.” Recall that Method (A) provides a percentage indicator of the accuracy of the translated sentence. In this method, the translation accuracy criteria vary depending on subjects. In contrast, Methods (B) and (C) provide a scale defined clearly in words. In these latter two methods, when the subjects had repaired a sentence to a certain level of accuracy, they may have finished the repair work.

In section 4.5, we showed that a mismatch between the subjective evaluation and the BLEU score may have occurred. The incorrect BLEU score may have affected subjects’ judgments.

We counted the number of inaccurate judgments that occurred when the BLEU score was incorrect. Table 7 shows the number of inaccurate judgments with an incorrect indication of the translation accuracy. From table 7, we see that approximately half of inaccurate judgments occurred when an incorrect indication was provided. We think that we can prevent the occurrence of these inaccurate judgments by improving the accuracy of the automatic evaluation method.

5.2 Appropriate Method for Accuracy Indication

In this section, we discuss evaluation point 2: “Which method is most effective?”

In section 4.4, we showed that subjects evaluated Method (B) as the best method for indicating the accuracy of a translation. From Table 6, Method (A) had the second smallest value of $\sum ar$, and the difference between the values for Methods (A) and (B) was small.

From Tables 5 and 6, Methods (A) and (B) received a high evaluation. However, in the free description section of the questionnaire, subjects provided the following comments: “Although Method (B) provides a clear scale, it is difficult to judge small differences,” “Method (A) makes it is easy to check variation of accuracy, but it is difficult to set evaluation standards.” From these comments, we found that a combination of Methods (A) and (B) was most effective, as each covered for the others shortcomings. Therefore, we think that a combination of these indicators may help users in judging translation accuracy more easily.

6 Conclusion

In this paper, we proposed methods to indicate the accuracy of a translation in order to encourage people to rectify inaccurate translations. We used an automatic evaluation method to measure the accuracy of the translated sentences by using the following three indicators: a percentage, a five-point scale, and a three-point scale. Moreover, we verified the effects of these indicators in reducing inaccurate judgments.

The results of the experiment were as follows.

(1) The indications of translation accuracy did not significantly affect the inaccurate judgments of users. However, the automatically evaluated values in this experiment were not always accurate. We think that the incorrect automatic-evaluated values may have led to some inaccurate judgments. If the accuracy of the automatic evaluation method improves, the translation accuracy indicators can help to reduce inaccurate judgments.

(2) The indication using a five-point scale obtained the highest evaluation and that using the percentage obtained the second highest evaluation. Further, the percentage indicator can cover for the shortcomings of the five-point scale. We believe that a combination of these indicators can help users in judging the translation accuracy more easily.

In the future, we will need to improve the accuracy of the automatic evaluation of the translation accuracy. Moreover, we will need to consider a method that reduces users' inaccurate judgment.

Acknowledgments. This work was partially supported by a Grant-in-Aid for Scientific Research (B), No. 22300044, 2010-2012.

References

1. Aiken, M.: Multilingual Communication in Electronic Meetings. ACM SIGGROUP, Bulletin 23(1), 18–19 (2002)
2. Tung, L.L., Quaddus, M.A.: Cultural differences explaining the differences in results in GSS: implications for the next decade. Decision Support Systems 33(2), 177–199 (2002)
3. Inaba, R.: Usability of Multilingual Communication Tools. In: Aykin, N. (ed.) HCII 2007. LNCS, vol. 4560, pp. 91–97. Springer, Heidelberg (2007)
4. Miyabe, M., Yoshino, T., Shigenobu, T.: Effects of Undertaking Translation Repair using Back Translation. In: Proceedings of the 2009 ACM International Workshop on Intercultural Collaboration (IWIC 2009), pp. 33–40 (2009)
5. Miyabe, M., Yoshino, T., Shigenobu, T.: Effects of Repair Support Agent for Accurate Multilingual Communication. In: Ho, T.-B., Zhou, Z.-H. (eds.) PRICAI 2008. LNCS (LNAI), vol. 5351, pp. 1022–1027. Springer, Heidelberg (2008)
6. Papineni, K., Roukos, S., Ward, T., Zhu, W.: BLEU: a Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 311–318 (2002)
7. NIST: Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics, Technical report, NIST (2002)
8. Uchimoto, K., Hayashida, N., Ishida, T., Isahara, H.: Automatic Rating of Machine Translatability. In: 10th Machine Translation Summit (MT Summit X), pp. 235–242 (2005)

9. Walker, K., Bamba, M., Miller, D., Ma, X., Cieri, C., Doddington, G.: Multiple-Translation Arabic (MTA) Part 1. Linguistic Data Consortium (LDC) catalog number LDC2003T18 and ISBN 1-58563-276-7
10. Ishida, T.: Language Grid: An Infrastructure for Intercultural Collaboration. In: IEEE/IPSJ Symposium on Applications and the Internet (SAINT 2006), pp. 96–100 (2006)
11. Kudo, T., Yamamoto, K., Matsumoto, Y.: Applying Conditional Random Fields to Japanese Morphological Analysis. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004), pp. 230–237 (2004)

Design of a Face-to-Face Multilingual Communication System for a Handheld Device in the Medical Field

Shun Ozaki¹, Takuo Matsunobe¹, Takashi Yoshino¹, and Aguri Shigeno²

¹ Faculty of Systems Engineering, Wakayama University,
930 Sakedani, Wakayama, Japan

² NPO Center for Multicultural Society Kyoto,
143 Manjuji-cho Shimogyo-ku, Kyoto, Japan
ozaki@yoslab.net, {matunobe,yoshino}@sys.wakayama-u.ac.jp,
aguri@tabunka-kyoto.org

Abstract. In the medical field, a serious problem exists with regard to communication between hospital staff and foreign patients. For example, medical translators cannot provide support in cases in which round-the-clock support is required during hospitalization. We propose the use of a multilingual communication support system called the Petit Translator between people speaking different languages in the hospital setting. From the results of experiments performed in such a setting, we found the following: (1) by clicking the conversation scene, the interface can retrieve the parallel text more efficiently than the paper media, and (2) when a questioner appropriately limits the type of reply for a respondent, prompt conversation can occur.

Keywords: parallel text, machine translation, handheld device, speech-to-speech translation, medical communication.

1 Introduction

The need for multilingual communication in Japan has increased due to an increase in the number of foreigners in the country. When people communicate in their nonnative language, the differences in language prevent a mutual understanding among the communicating individuals [1, 2]. In the medical field, this can create a serious problem when it comes to communication between hospital staff and patients. Currently, medical translators accompany patients to medical care facilities, and the number of requests for such translators is increasing. Medical translators cannot provide support at all times, however, especially in cases where round-the-clock support is required or in the case of hospitalization. Hence, a system that supports accurate multilingual communication is required.

We have developed a support system for multilingual medical reception termed M³ [3]. In this study, we cover the hospital setting, in which members of the medical staff and foreign patients communicate with each other in various places throughout the hospital. We feel that such a setting calls for a translation system. We propose the use of a multilingual communication support system, called the Petit Translator, between people speaking different languages during hospitalization. This system uses parallel

texts for accurate multilingual communication and machine translation for daily conversation. In this study, we describe the development of the Petit Translator and offer an evaluation of the system.

2 Related Works

Parallel texts are lines of text in 1 language paired with translations of that text in another language. In other words, parallel texts are accurate translations pre-paired in advance that are meant to improve the efficiency and accuracy of medical treatment [4]. Face-to-face communication systems using parallel texts are now in use. One of these is a support system using speech-to-speech translation for foreign travelers [5]. Another topic of research is a tool that supports communication between speakers of different languages and uses parallel texts for speech recognition [6]. In systems such as this, the user inputs speech and the system outputs a translated sentence; however, the system cannot output sentences that have not been previously registered. Speech translation systems using phrase translation for communication in the medical field have also been proposed [7, 8]. Even these systems, however, provide insufficient support for the medical field.

3 Petit Translator

The Petit Translator supports 5 languages: Japanese, English, Chinese, Korean, and Portuguese. It can operate on Android Devices (a smart phone). In the following, we present the functions of the Petit Translator.

3.1 Voice Translation Function

The speech translation function is one that translates the spoken word into another language, and creates the sound and characters. Figure 1 shows the system configuration of the Petit Translator.

1. Voice input function

The voice input function converts the voice into characters. We use a voice input function because it is difficult for users to input characters manually on a small screen. We use Google voice recognition for the voice input function.

2. Translation function (Similar parallel text retrieval and machine translation)

The translation function retrieves the parallel text and carries the machine translation concurrently. Figure 2 shows the result of the translation. The system shows the result of the parallel text retrieval and machine translation on the same screen (Figs. 2 (2) and (3)). A user can use the machine translated sentence when no results have been found by the system. The system shows the back-translated sentence to check the accuracy of the machine translation (Fig. 2 (3)). We use the Web service Language Grid [9] to retrieve the parallels texts.

3. Voice synthesis function

The voice synthesis function synthesizes the voice from the parallel text or result of the machine translation. We use the voice synthesis service of Language Grid. In the

hospital, there are places where precision equipment is prohibited. Therefore, it is difficult to communicate through characters or by using a screen. In these cases, a medical staff member can communicate with a patient using the voice synthesis function.

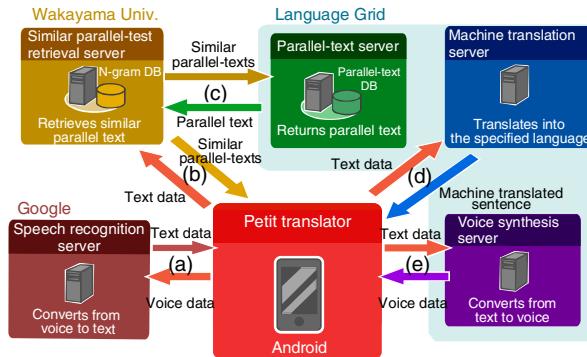


Fig. 1. System configuration of the Petit Translator

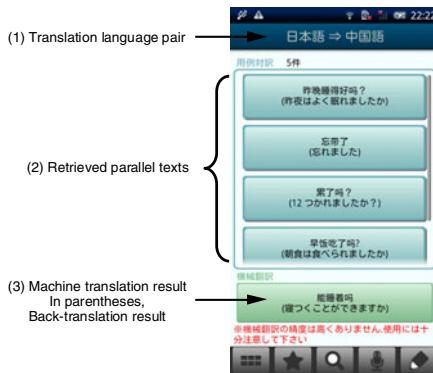


Fig. 2. Screenshot of the result of translation

3.2 Response Function

The response function is a function that allows the patients to respond when the system is shown. Figure 3 shows a screenshot of the response function.

1. Function for a questioner

When a medical staff member selects the translation result, that result is shown on the screen for the respondent. He or she can show the screen and synthesize the voice (Fig 3 (3)).

2. Function for a respondent

A respondent can answer the question using the “yes” or “no” buttons (Fig. 3 (4)). When a respondent needs to answer in detail, he or she can use the machine translation to input text for a detailed response (Fig. 3 (5)).

3.3 Scene Retrieval Function

The scene retrieval function is a function to make a short list of the parallel texts by selecting the conversation scene. The time of the input can be saved by clicking the conversation scene.



Fig. 3. Screenshot of a response screen

4 Experiment

4.1 Purpose of the Experiment

The purpose of the experiment was to verify the efficiency of the Petit Translator. We compared a case in which the Petit Translator was used with a case in which a multilingual leaflet and an electronic dictionary were used between a medical staff member and a foreign patient. We used a multilingual leaflet that was provided by the Mitsubishi Tanabe Pharma Corporation. We compared both the efficiency and accuracy.

4.2 Subjects and Procedure

A total of 18 subjects (9 Japanese and 9 Chinese) participated in the study. All were students at Wakayama University. One Japanese subject and 1 Chinese subject participated in each experiment. There were 9 experiments altogether. None of the Japanese subjects had a smart phone, but 3 of the Chinese subjects had one.

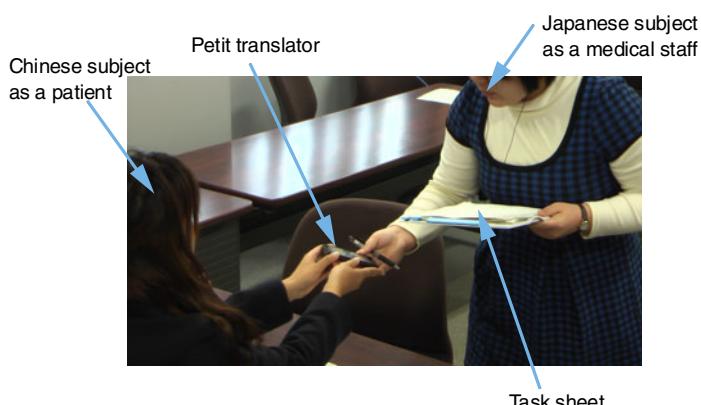
All of the Chinese subjects could speak Japanese and read simple Japanese in daily life. The experiment assumes that the conversation is between users who do not understand each other. Therefore, we prohibited using Japanese in the experiments. Both subjects wore earphones to prevent hearing a Japanese voice.

4.3 Tasks of the Experiment

We allocated the Japanese subjects to the role of medical staff member and Chinese subjects to the role of patient. In the experiments, which used each system, there were 5 tasks respectively. Table 1 shows the contents of each task. The task was initiated

Table 1. Tasks of the experiment

Task		Reply method		
	Japanese subject (as a medical staff member)	Chinese subject (as a patient)	Petit Translator	A multilingual leaflet and an electronic dictionary
(1)	Please confirm whether the patient ate breakfast.	Answer the question.	Medical staff member: Ask the question using the scene retrieval function. Patient: Answer using the “Yes” or “No” buttons.	
(2)	Please tell the patient to take 2 types of medicine before he or she sleeps.	Description of the instruction	Medical staff member: Give instructions using the scene retrieval function. Patient: Write down the instructions.	
(3)	Please ask the patient to confirm that he or she is worried about something.	Please answer that you are worrying about hospital fees.	Medical staff member: Ask the question using the scene retrieval function. Patient: Answer using a text area for detailed response (machine translation).	They communicate with a medical leaflet and an electronic dictionary only.
(4)	Please ask the patient to confirm whether there is pain somewhere.	Please answer that you have a stomachache with a pricking sensation.	Medical staff member: Ask the question using the general retrieval function. Patient: Answer using a text area for detailed response (machine translation).	
(5)	Please ask the patient to describe a symptom that he or she has suffered and the period he or she suffered it.	Please describe the symptom you suffered and the period you suffered it.	Medical staff member: Ask a question using the machine translation. Patient: Answer using a text area for detailed response (machine translation).	

**Fig. 4.** A photograph of an experiment using the Petit Translator

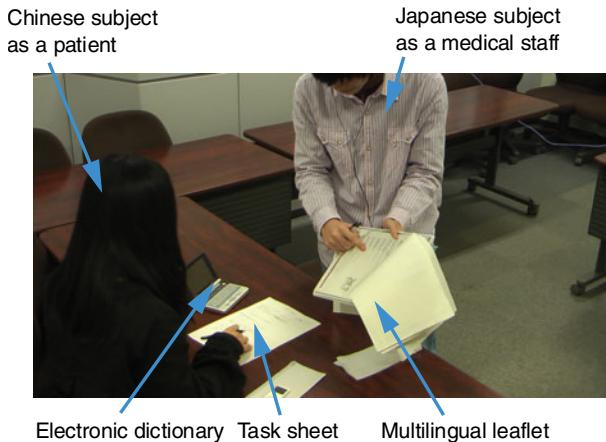


Fig. 5. A photograph of an experiment using a multilingual leaflet and an electronic dictionary

by a Japanese subject. Each task was completed when the Chinese subject had responded to the Japanese subject's question. When the 5 tasks were completed, the system (the Petit Translator or a multilingual leaflet and an electronic dictionary) was switched, and the 5 tasks were performed again. Figure 4 shows a photograph of an experiment using the Petit Translator. Figure 5 shows a photograph of an experiment using a multilingual leaflet and an electronic dictionary. Before the experiments, the subjects practiced the operation of the Petit Translator and the electronic dictionary. They also browsed the multilingual leaflet.

5 Discussion

5.1 Time to Accomplish Tasks

Table 2 shows the time required by the Japanese and Chinese subjects to accomplish the tasks. In order to discuss the time required to accomplish the tasks, we have divided the Japanese subjects from the Chinese subjects.

1. Time required by the Japanese subjects to accomplish the tasks

The Petit Translator can accomplish tasks (1) and (2) in less time than a multilingual leaflet and an electronic dictionary. With tasks (4) and (5), however, the Petit Translator requires more time than a multilingual leaflet and an electronic dictionary. When using the Petit Translator with tasks (4) and (5), the subject needs to input a retrieval word. The subjects took a long time to do this because they were inexperienced in inputting using a software keyboard.

Three of the 9 Japanese subjects used the voice input. Three used the voice input only once because they did not get a good result. The other 6 subjects answered, "I don't hear my voice easily because I used earphones for the experiment. So, I didn't use the voice input."

2. Time required by the Chinese subjects to accomplish the tasks

The Petit Translator requires more time than a multilingual leaflet and an electronic dictionary to accomplish tasks (3), (4), and (5). We found that the user interface of the Petit Translator has some problems. Moreover, the Chinese subjects have experience in using an electronic dictionary, and use one often. Therefore, the time required to accomplish the tasks is shorter when using the multilingual leaflet and the electronic dictionary than when using the Petit Translator.

Table 2. Time required by the Japanese and Chinese subjects to accomplish the tasks

		Japanese subject (as a medical staff member)		Chinese subject (as a patient)	
Task	Type	Average (sec)	SD (sec)	Average (sec)	SD (sec)
(1)	Petit	35	44	<1	<1
	Leaflet	67	48	<1	<1
(2)	Petit	54	51	-	-
	Leaflet	102	47	-	-
(3)	Petit	63	38	137	198
	Leaflet	53	21	34	20
(4)	Petit	116	62	85	47
	Leaflet	66	75	33	24
(5)	Petit	236	73	162	120
	Leaflet	170	88	81	35

- SD means standard deviation.

- Task (2) for the Chinese subjects was to write down the instructions.

- Petit stands for the Petit Translator.

- Leaflet stands for a multilingual leaflet and an electronic dictionary.

5.2 Efficiency of Conversation

After the experiment, we asked the subjects to fill out a questionnaire. We used a 5-point Likert scale (1: strongly disagree, 2: disagree, 3: neutral, 4: agree, and 5: strongly agree) and a free description for the evaluation. Table 3 shows the results of the questionnaires of the Japanese subjects who acted as medical staff members. We found that they were able to tell their intentions to the patients quickly using the Petit Translator (Table 3 (1)). In the free description, we obtained the comment, “I was able to present the content to the patient with sentences by using the Petit Translator.”

Table 4 shows the results of the questionnaires of the Chinese subjects who acted as patients. However, we found that no significant difference was seen with the Chinese subjects (Table 4 (1)).

Table 5 shows the results of the questionnaires regarding searching for parallel texts. We found that using the Petit translator is easier than using a multilingual leaflet and an electronic dictionary in searching for parallel texts (Table 5).

In the free description, we obtained the comment, “When I used the electronic dictionary, it was necessary to look up some of the words, which was troublesome” and “It takes time to look for a parallel text with shuffling through the multilingual leaflet.” Regarding the scene retrieval function, we obtained the comment, “It was easy to have looked for because it was classified by the scene.”

5.3 Accuracy of Conversation

We found that the Chinese subjects could understand the intentions of the medical staff easily (Table 4 (2)).

Some Japanese subjects guided the Chinese subject to receive the expected response from him or her. For example, they had passed the Petit translator to the Chinese subject with the keyboard had been displayed when a detailed response was expected. In the free description, we obtained the positive comment, “Only 1 item came out on the screen, and I answered it easily.” The Japanese subjects understood the patients’ intentions well using both systems (Table 3 (2)).

Table 3. Results of the questionnaires of the Japanese subjects as medical staff

Question	Type	Evaluation (people)					Median
		1	2	3	4	5	
(1) I was able to express my intention to the patient quickly.	Petit	0	3	1	5	0	4
	Leaflet	2	4	1	2	0	2
(2) I quickly understood the intention that the patient expressed.	Petit	1	0	2	2	4	4
	Leaflet	0	3	1	5	0	4

- Petit stands for the Petit Translator.

- Leaflet stands for a multilingual leaflet and an electronic dictionary.

Table 4. Results of the questionnaires of the Chinese subjects as patients

Question	Type	Evaluation (people)					Median
		1	2	3	4	5	
(1) I was able to express my intention to the medical staff member quickly.	Petit	0	2	4	2	1	3
	Leaflet	0	4	4	0	1	3
(2) I quickly understood the intention that the medical staff member expressed.	Petit	0	1	0	5	3	4
	Leaflet	0	3	4	0	2	3

- Petit stands for the Petit Translator.

- Leaflet stands for a multilingual leaflet and an electronic dictionary.

Table 5. Results of the questionnaires regarding searching for parallel texts

Question		Evaluation (people)					Median
		1	2	3	4	5	
(1) It was easy to look for the parallel texts by using the Petit translator.		0	1	1	2	4	4.5
(2) It was easy to look for the parallel texts by using a multilingual leaflet and an electronic dictionary.		4	4	1	0	0	2

- One person did not respond to question (1), so the answer total for question (1) was 8 people.

6 Conclusion

We have developed a multilingual communication support system called the Petit Translator for use between people speaking different languages in the hospital setting. There are 3 main cases in which translation is needed in the hospital setting: (1) for use throughout the hospital, (2) when a requested conversation requires accuracy, and (3) when quick correspondence is necessary. Therefore, the system provides 3 types of features: (1) portability, (2) the combined use of parallel texts and machine translation, and (3) the use of easy input methods for word retrieval (voice input and button click input).

From the results of experiments done in the hospital setting, we obtained the following results:

- (1) By clicking the conversation scene, the interface can retrieve the parallel text more efficiently than the paper media.
- (2) When a questioner appropriately limits the type of reply for a respondent, prompt conversation can occur.

We are planning to introduce the Petit Translator to some hospitals in the near future.

Acknowledgment. This work was partially supported by the Strategic Information and Communications R&D Promotion Programme (SCOPE) of the Ministry of Internal Affairs and Communications of Japan.

References

1. Aiken, M.: Multilingual communication in electronic meetings. *ACM SIGGROUP Bulletin* 23(1), 18–19 (2002)
2. Tung, L.L., Quaddus, M.A.: Cultural differences explaining the differences in results in gss: implications for the next decade. *Decision Support Systems* 33(2), 177–199 (2002)
3. Miyabe, M., Fujii, K., Shigenobu, T., Yoshino, T.: Parallel-text Based Support System for Intercultural Communication at Medical Receptions. In: Ishida, T., R. Fussell, S., T. J. M. Vossen, P. (eds.) *IWIC 2007. LNCS*, vol. 4568, pp. 182–192. Springer, Heidelberg (2007)
4. Wang, B., Cheng, X., Bai, S.: Example-based phrase translation in Chinese-English CLIR. In: The 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 435–436 (2002)
5. Ikeda, T., Ando, S., Satoh, K., Okumura, A., Watanabe, T.: Automatic interpretation system integrating free-style sentence translation and parallel text based translation. In: The Workshop on Speech-to-Speech Translation: Algorithms and Systems, pp. 85–92 (2002)
6. Imoto, K., Sasajima, M., Shimomori, T., Yamanaka, N.: A multimodal supporting tool for multi lingual communication by inducing partner's reply. In: The 11th International Conference on Intelligent User Interfaces (IUI 2006), pp. 30–332 (2006)
7. Rayner, M., Bouillon, P., Dalsem, V.V., Isahara, H., Kanzaki, K., Hockey, B.A.: A limited-domain English to Japanese medical speech translator built using regulus 2. In: The 41st Annual Meeting on Association for Computational Linguistics 2 (ACL 2003), pp. 137–140 (2003)
8. Chung, J.W., Kern, R., Lieberman, H.: Topic spotting common sense translation assistant. In: CHI 2005 Extended Abstracts on Human Factors in Computing Systems, pp. 1280–1283 (2005)
9. Ishida, T.: Language grid: an infrastructure for intercultural collaboration. In: IEEE/IPSJ Symposium on Applications and the Internet (SAINT 2006), pp. 96–100 (2006)

Computer Assistance in Bilingual Task-Oriented Human-Human Dialogues

Sven Schmeier, Matthias Rebel, and Renlong Ai

Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI)

Alt-Moabit 91c, 10559 Berlin, Germany

{schmeier, matthias.rebel, renlong.ai}@dfki.de

Abstract. In 2008, the percentage of people with a migration background in Germany had already reached more than 15% (12 Million people). Among that 15%, the ratio of seniors aged 50 years or older was 30% [1]. In most cases, their competence of the German language is adequate for dealing with everyday situations. However sometimes in emergency or medical situations, their knowledge of German is not sufficient to communicate with medical professionals and vice versa. These seniors are part of the main target group within the German Ministry of Research and Education (BMBF) research project SmartSenior [2] and we have developed a software system that assists multilingual doctor-patient conversations to overcome language and cultural barriers. The main requirements of such a system are robustness, accurate translations in respect to context and mobility, adaptability to new languages and topics and of course an appropriate user interface. Furthermore, we have equipped the system with additional information to convey cultural facts about different countries. In this paper, we present the architecture and ideas behind the system as a whole as well as related work in the area of computer aided translation and a first evaluation of the system.

Keywords: language barriers, human-human dialogue system, health care.

1 Introduction

The research project SmartSenior aims to develop technological and comprehensive solutions to aid senior citizens in leading independent lives. The aim is to support elderly people in their day-to-day lives and social interactions, and, in terms of health, support them in a way that helps them continue living within their familiar environment. The project targets both senior citizens leading overall independent lives, as well as acutely or chronically ill elderly people in need of assistance and care.

Within the context of “support”, a subproject in SmartSenior is dedicated to helping seniors with migration backgrounds carry out conversations, especially in emergency situations.

High quality patient-centered care depends heavily on communication. Understanding the patient's needs and abilities to follow their doctor's advice are the keys for successful treatment [3]. In multicultural, multilingual contexts, language and cultural barriers present a key challenge to effective communication between patient

and service provider. Unfortunately, access to translation services or human translators, is not always available, especially in emergency situations where time is a very limiting factor. Numerous studies show the effect that the absence of language interpreters have in such situations ([4], [5], and [6]). Furthermore, those translators need to be trained to handle specific situations, especially when dealing with patients that come from different social cultures. It is also known that interpreter errors may occur if untrained ad-hoc interpreters are used [7].

Starting from these observations, we developed a software system that makes conversations across different languages possible.

The main requirements on such a system are listed in the following points:

1. Robustness: the system should work robustly and not quit services under any circumstances
2. Accuracy: the system should provide accurate translation
3. Mobility: the system must be accessible anywhere and anytime
4. Adaptivity: the system should be able to handle new languages and topics
5. Usability: the user interface must be designed with respect to the target group

To meet these requirements the following design, implementation and platform decisions have been made: with respect to mobility and usability, we chose touch-sensitive Tablet PC as our target platform. The software system is divided into UI and database content sections. This facilitates the easy incorporation of new languages, situations and information. The system itself is designed as a cross lingual, mixed-initiative dialogue and information system. The content of the dialogues have been created with the help of doctors and medical staff personnel with respect to the given task. The translations have been done by human expert interpreters. The system itself runs off-line, no internet connection is needed.

In the remainder of this paper, we first present related work that has been done in the area of automated translation (services). In the next sections, we describe the system specifications, the dialogue design and the evaluation of our first in house test series. We finally end with the conclusion and outlook.

2 Related Work

Overcoming language barriers has been the focus of research for more than six decades. Since then much progress has been made, starting from bilingual dictionaries and hand coded rules for the final translation, to linguistic parsing of the input in one language generating the output in a second language. However, due to the high ambiguity of natural language the results are still poor compared to the costs. Restrictions of the domain improve translation quality, but the process of adapting new or different domains turned out to be very costly.

With the fast developing and growing internet and search engine companies, huge amounts of corpora have been collected in the last decade. Hence research has focused more and more on pure statistical machine translation. Machine translation services like Babelfish¹ or Google Translate² are now available for the public. Unfortunately,

¹ <http://babelfish.yahoo.com/>

² <http://translate.google.com>

the accuracy is still quite poor. For example, the German sentence: “Haben Sie heute morgen schon Wasser gelassen?” (Engl.: “Did you urinate this morning?”) produces the translation: “Did you this morning has left the water?”

Another kind of translation service are phrasebooks which contain frequently used sentences and their translation. In the last two years, more and more electronic phrasebooks have become available for mobile devices. Phrasebooks allow users to express a limited set of utterances in a language they do not know. Unfortunately, the partner is unable to reply, i.e. they do not allow dialogues at all. However, phrasebook translation is more accurate because the phrases have been translated by humans.

In the research project COMPASS2008 [8] a bilingual information and translation service system for mobile devices was developed. The goal of the project was to develop a system to empower foreign visitors during the 2008 Olympic Games that integrates functionalities of a phrasebook, a translation aid and a powerful information system that is connected to various services via the internet [9]. The main ideas behind this system have directed our research for overcoming language and cultural barriers.

3 System Specification

Our cross lingual translation and information system consists of two main components: One component is the system for overcoming language and cultural barriers that runs on the user's tablet. Dialogues and information data are stored and maintained in a database and our user interface as well as the Text-To-Speech (TTS) system also run on the device. Its second component is the system for creating and managing dialogues, cultural information and translations.

3.1 System on the Device

Our system is designed according to the Model View Controller (MVC) pattern [10]. The user interface, the database containing the cross lingual dialogues and dictionary, and the interaction controller between these two layers form independent modules which communicate via object messages. It runs as an App on an Apple iPad with the underlying operating system iOS 3.2 and does not need an internet connection.

3.2 Functionality

We implemented our solution as a mixed-initiative dialogue system. The dialogues are divided into several main dialogue situations. Each situation provides several phrases and a set of more specific sub-situations, which again contain several phrases. The user initiates a dialogue by selecting a phrase in his language. This phrase is then translated and presented to the partner along with a set of phrases designed as responses. The partner may then select a phrase which in turn is translated and presented to the user along with a relevant set of possible responses and so on.

The system provides three types of phrases:

1. **Static phrases** are sentences that are translated immediately after being selected/touched.
2. **Dynamic phrases** contain slots that have to be completed. These slots have a specific type with expected values. Possible types are (a) part_of_body, (b)

diseases, (c) allergies, (d) dictionary and so on. They are represented as clickable buttons within the phrases. There are four distinct types of slots: select lists (like a, b, c), empty fields to fill with values (i.e. measuring units) and special forms to enter time or date expressions, and selectable pictures to enhance communication and understanding.

3. **Information phrases** can either be static or dynamic and they contain an additional information symbol. By touching this symbol the system provides context-aware cultural information.

The system also takes into account that sometimes patients may not be able to respond by using the device. In these cases, the doctor is able to use an appropriate single mode with further options (i.e. phrases like “Wink, if you can hear me!”, “Wink once for no and twice for yes!”, “Everything will be fine, I’ll take care of you.”).

Fig.1 shows an example dialogue between a Turkish speaking senior and a German speaking doctor. The doctor first chooses the desired dialogue situation, in our case “Notfall zu Hause” (Emergency At Home) shown on the left screen. After that the system computes the phrases that will be displayed to the doctor. In our case, the doctor chooses the phrase “Können Sie beschreiben was genau passiert ist?” (“Please describe what happened.”) which will be translated for the senior as “Tam olarak ne olduğunu anlatabilir misiniz?”. Now the senior can answer the doctor by simply touching one of the phrases offered.

The second screen displays the information phrase “Ich muss Sie jetzt untersuchen.” (Engl.: “I need to examine you”). After clicking on the information symbol, the system shows the following hint: A western male doctor should not examine female persons with a Muslim background without any relatives in the room.



Fig. 1. Screens of the dialogue system on the mobile device

3.3 Additional Components on the Device

The system architecture allows the integration of additional components. So far we have experimented with two components in order to provide alternative, more comfortable and less restricted forms of user interaction: Text-To-Speech (TTS) synthesis and Automatic Speech Recognition (ASR).

Speech-Output – TTS. In many situations, it is important for the system to provide not only visual output but also spoken language - especially in situations where the senior is not able to read, e.g. because of impaired vision or illiteracy. Some languages, for example, Chinese, use special symbols for certain technical terms which may be unknown to the senior but are understandable when spoken. Spoken language also helps increase confidence between the doctor and patient.

The fact that our architecture utilizes dynamic phrases makes it impossible to prerecord the phrases and play them when necessary. A phrase like “Please give me [dictionary]” could produce thousands of translations depending on the user’s input for the slot dictionary. That is why we decided to make use of the commercial TTS system products already available on mobile devices. We tried three vendors: Loquendo, Nuance and SVOX. All systems provide voice output, synthesized spoken language, which sounds natural and is easy to understand. However for technical reasons, we chose the Mobile TTS Standard by SVOX [11]. It supports 27 languages including nearly all European languages such as: German, English, French, Spanish, Italian, Danish, Turkish, etc. It also supports languages like Mandarin, Cantonese, Arabic, Taiwanese, and Korean. Further key features that are important for our system are:

- Dynamic switching between languages
- Flexible voice control in terms of speed, pitch and volume: for our target group, the voice should be loud, slow and have a clear pronunciation
- Support of phonetic alphabets: this feature guarantees correct articulation for names of medical products
- Synthesis of mixed languages: this is useful for the names of specific products or people

Speech Input – ASR. At the moment, it is not technologically possible to perform 100% accurate speech recognition and on portable devices where CPU and memory are restricted, the ability to handle dictation (i.e. to recognize everything spoken) is even less reliable. However if the number of phrases that have to be recognized are limited, recognition accuracy can be improved dramatically. In our case, although we do provide phrases that cover many situations and their sub-situations, the number of these phrases is still limited. Under this condition, an ASR system that recognizes these phrases can be deployed to our target portable device. We have run tests with Nuance, SVOX and Fonix ASR engines on HTC HD mobile phones³ (528MHz CPU and 288 MB RAM) in both English and German. The best engine reached 72.69% accuracy for recognition. This result relates to a 68.67% recognition rate for outdoor usage and 74.28% for indoor usage. [12].

Taking into account that our tests were carried out by colleagues in a stress-free situation who were willing to speak very clearly and who held the device at the proper distance and in the right direction, we estimate that the result will decrease significantly in real life emergency situations with seniors. So for the moment, we have abandoned the idea of implementing ASR in our final system. In the future we plan to continue testing and evaluating ASR systems on the iPad. (See last chapter for more information).

³ At the time of our tests, the ASR systems running on the iPad were still under development.

3.4 Dialogue Management System (DMS)

The system's quality depends heavily on the dialogues and information that are provided. This pertains to both the quantity of the dialogue situations and their sub-situations as well as to the quality of the questions and their answers. Each emergency situation is different in respect to the course of events but also with respect to the people involved and though the dialogues should be as general as possible, it is important that each conversation seem unique to the patient. A further factor is the screen size which plays a very limiting role for the development of the dialogues. The sentences or phrases can not be too long but they have to include the main information needed for the dialogue.

Our DMS has been designed as a browser-based WYSIWYG⁴ system that fully supports the dialogue's author, editor and translator in their specific needs. It provides all necessary operations, such as:

- adding and deleting languages
- translating phrases, dictionary entries or names of dialogues and sub-dialogues
- as well as adding/deleting/modifying dialogues, sub-dialogues, phrases and slot types and content

All of these different levels - i.e. representation of the system, the dialogue structure and the phrases in their context - are important for developing advanced dialogues and providing context-aware translations. That is why we provide two views for the managing environment. One is an application like view (Fig 2, left) and the other one is a ScalableVectorGraphic⁵ representation (Fig.2, right) of a specific dialogue situation which gives a general overview of context, structure and flow. The DMS exports the dialogues, which were created, into the appropriate database format for the system on the device.



Fig. 2. Screens of the DMS with described functionalities

4 Task-Oriented Structure of Dialogues

In this section, we introduce in more detail some of the dialogues that have been implemented in our system and the rationale behind their development. There are many

⁴ What You See Is What You Get.

⁵ SVG support to augment objects in the picture with URLs.

different places and situations in which medical assistance is carried out for e.g. at hospitals, at home, during transport etc. Each of these situations are represented as a main menu item. Hence at the first level, there are four Situations: *Emergency at Home*, *Emergency On the Way*, *Ambulance* and the situation independent topic *Anamnesis*.

Each situation provides, with respect to its topic, a set of specific phrases and sub-situations with specific phrases. In case of an accident, it is important to understand what happened and why in order to better understand which injuries or problems were caused by the accident. The answers to these question lead to the doctor's first assessment. In severe health incidents, the doctor needs to act immediately making it necessary that the number of questions are limited and well organized. They also have to be simple and easy to understand and each question should be answerable by yes or no. In extreme cases, the patient may be immobilized and unable to operate the device, so we provide questions that allow alternative ways for the user to respond within the dialogue (like nodding or head shaking or blinking). There are also situations, when no information is expected or the patient can not respond at all. These dialogues are designed to calm down the patient and explain what has happened and is happening. In such situations, the capability of the system to produce TTS synthesis is especially important.

The area of medical examinations is a wide field and exhaustive coverage can clearly not be expected from the limited dialogues provided in our system. Especially the anamnesis of a patient is very difficult and tedious to model. Nevertheless, the British Red Cross offers an emergency phrasebook⁶ which covers the most common medical questions and terms to help first-contact-staff communicate with patients who do not speak English. This phrasebook is meant to make initial assessment in case there is no interpreter. For our system, we drew on the phrases contained in the phrasebook and added specialized sentences for seniors.

5 Evaluation

The system will be fully evaluated in field-tests in November 2011. Currently, we have finished the first test series. A second test series specifically to evaluate the UI design for elderly people is planned.

5.1 Setup

Each test was performed by two persons⁷ role-playing a doctor-patient dialogue. After a brief introduction to the system and how to use it, the testers were asked to perform three different tasks (easy, medium and hard). The order of the tasks was randomized for each pair of testers. After each task, both testers had to rate several statements on a Likert scale and a general questionnaire had to be filled out after completing the entire test with questions concerning the user interface, dialogue structure, and difficulties encountered during the course of the dialogue etc.

⁶ http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_4073230

⁷ The testers had not seen the system before.

5.2 Results

Most people judged the system as an appropriate means of communication between people who speak different languages (58% agree, 14% totally agree). More than 70% of the testers assessed the level of politeness and naturalness of the presented phrases as good. The rating for the dialogue flow, i.e. dialogue structure and hierarchies, the efficiency of the phrases, and the time needed to respond was mixed. We observed that subdividing a dialogue in sub-dialogues sometimes and too many phrases on one page produce longer latencies per turn. Furthermore, the testers had some problems with dynamic phrases, i.e. phrases with slots that had to be completed. We noticed that, independent of task order, these problems became less prominent during the course of the test.

In summary, we gained useful feedback concerning the user interface as well as dialogue and phrase structures. The key issues to improve the system's usability will be determining the optimal weight between static vs. dynamic phrases and the ratio between dialogue hierarchy, number of phrases per page and general expressivity.

6 Conclusion and Outlook

We presented a system that helps overcome language and cultural barriers between doctors or medical staff and seniors with a migration background. The system is designed as a cross lingual, mixed-initiative dialogue and information system running on a Tablet PC. These conditions meet our main requirements: robustness, accurate translation, mobility, scalability with respect to new languages and content as well as appropriate user interfaces for seniors.

The content for the dialogues and information have been designed and developed with the help of doctors and medical staff. We have also built a Dialog Management System which allows us to easily integrate content into the system.

A first in house evaluation has been done which confirmed several aspects concerning the general user interface, the appropriateness of the dialogues and the overall handling of the various system components including the Text-To-Speech module.

From field-tests planned for November 2011 and the beginning of 2012, we expect to gain new insights and results that will be integrated in our system. Even though our dialogues and content are not yet certified for hospitals, we hope our system can be a beginning to help support the integration of people with a migration background into society.

In the near future, we will concentrate on incorporating speech input into our system. As we learned from an interview with a practitioner the capability of recognizing speech would be extremely useful for patients and doctors. Currently, we are evaluating Speech Recognition (ASR) systems on the iPad and hope to find ways to achieve satisfactory recognition rates. We also have to evaluate whether the speed, i.e. the time between speaking and retrieving the results is reasonable in this context. Furthermore, the current rapid development of tablets will no doubt have positive effects on developmental concerns. We expect a vast improvement of hardware and

operating systems which will lead to better noise reduction for outdoor usage, better signals for the ASR module and the possibility of accessing larger speech models for improved recognition accuracy.

Acknowledgements. This system has been developed in the project SmartSenior. The SmartSenior project consists of a consortium of 29 project partners of either industrial or research back-ground and it is funded by the German Ministry of Research and Education (BMBF, FKZ 16KT0902) within the course of the “High-Tech-Strategie-Deutschland” program.

References

1. Statistisches Bundesamt: Bevölkerung und Erwerbstätigkeit; Bevölkerung mit Migrationshintergrund - Ergebnisse des Mikrozensus 2008. Wiesbaden (2010)
2. SmartSenior (FKZ 16KT0902), <http://www.smart-senior.de>
3. Elderkin-Thompson, V., Silver, R.C., Waitzkin, H.: When nurses double as interpreters: a study of Spanish-speaking patients in a US primary care setting
4. Siejo, R., Gomez, H., Friendenberg, J.: Language as a communication barrier in medical care for Hispanic patients. In: Padilla, A.M. (ed.) Hispanic psychology-critical issues in theory and research, pp. 169–181. Sage Publication, California (1995)
5. Smedley, B.D., Stith, A.Y., Nelson, A.R.: Unequal treatment. Confronting racial and ethnic disparities in health care. The National Academies Press, Washington (2003)
6. Flores, G.: The impact of medical interpreter services on the quality of health care: a systematic review. *Med. Care Res. Rev.* 62(3), 255–299 (2005)
7. Flores, G., Laws, M.B., Mayo, S.J., Zuckerman, B., Abreu, M., Medina, L., Hardt, E.J.: Errors in medical interpretation and their potential clinical consequences in pediatric encounters. *Pediatrics* 111(1), 6–14 (2003)
8. Aslan, I., Xu, F., Uszkoreit, H., Krüger, A., Steffen, J.: COMPASS2008: Multimodal, multilingual and crosslingual interaction for mobile tourist guide applications. In: Maybury, M., Stock, O., Wahlster, W. (eds.) INTETAIN 2005. LNCS (LNAI), vol. 3814, pp. 3–12. Springer, Heidelberg (2005)
9. Uszkoreit, H., Xu, F., Liu, W., Steffen, J., Aslan, I., Liu, J., Müller, C., Holtkamp, B., Wojciechowski, M.: A successful field test of a mobile and multilingual information service system COMPASS2008. In: Jacko, J.A. (ed.) HCI 2007. LNCS, vol. 4553, pp. 1047–1056. Springer, Heidelberg (2007)
10. Gamma, E., Helm, R., Johnson, R., Vlissides, J.: Design Patterns: Elements of Reusable Object-Oriented Software. Addison-Wesley Longman, Amsterdam (2004)
11. SVOX AG: SVOX Manual: Speech Output Engine SDK 4.3.1, July 17 (2009)
12. Ai, R.: Fuzzy Match between Speech Input and Textual Database, Master’s thesis, Technical University Berlin (2010)

Developing and Exploiting a Multilingual Grammar for Human-Computer Interaction

Xian Zhang, Rico Andrich, and Dietmar Rösner

Department of Knowledge Processing and Language Engineering,
Otto-von-Guericke-University Magdeburg,
P.O. Box 4120, D-39106 Magdeburg, Germany
{xzhang,roesner}@iws.cs.uni-magdeburg.de,
ricoandrich@gmx.de

Abstract. How to build a grammar that can accept as many as possible user inputs is one of the central issues in human-computer interaction. In this paper, we report about a corpus-based multilingual grammar, which has the aim to parse naturally occurring utterances that are used frequently by subjects in a domain-specific spoken dialogue system. The goal is achieved by the following approach: utterance classification, syntax analysis, and grammar formulation.

Keywords: NLU, HCI, grammar, multilinguality, GF.

1 Introduction

In human-computer interaction (HCI), the problem of grammatical coverage, which means how to build a grammar that can accept as many as possible user inputs, is one of the central issues in spoken dialogue systems.

In this work we are thus interested in the problem of developing a grammar that is powerful enough to parse as many as possible naturally occurring utterances that are frequently used by subjects in a domain-specific spoken dialogue system. The idea is to first classify user utterances in order to detect the commonly used patterns and concepts from a corpus, and then find out the syntactic structures and summarize them into rules to build up the grammar.

The development was based on the transcripts of the NIMITEK corpus [1] that collects recordings of affected German speech in human-computer interaction (HCI) gathered in a Wizard-of-Oz (WOZ) [2] experiment that was conducted by Gnjatovic and Rösner [3] since 2006. The NIMITEK corpus is a multimodal corpus of affected behavior in human-machine interaction [4]. It contains 15 hours of audio and video recordings that are produced during the refined WOZ experiment, which was designed to induce emotional reactions of the human participants. During the experiment, the language of the subjects was very natural because the subjects were only allowed to give speech based instructions to the system and no language restrictions were given to them.

In this work, we use one of the tasks in the NIMITEK corpus named TANGRAM as the sample. Tangram is a popular match game. Of seven stones, namely, five triangles, a square and a parallelogram, we can put figures together. All pieces must

be used. They can be moved, rotated, or reflected; they have to touch each other, but must not overlap. Fig. 1 shows the seven stones that are used in Tangram, and an example of Tangram figure.

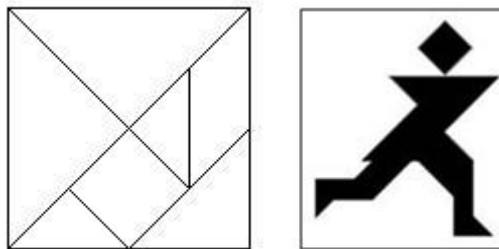


Fig. 1. Tangram stones and an example of Tangram figure

To illustrate how natural utterances of the subjects in solving of Tangram puzzles look like, let us observe two dialogue fragments translated into English from the corpus shown in Fig. 2. The example includes two sequences of commands produced by the subjects, whereas the corresponding system's actions are not explicitly given.

In the first sequence, from the subject's point of view, the square should be put on the position of the head in the figure (see Fig. 1), but unfortunately the machine, i.e. WOZ pretended that it did not understand this natural speech (e.g. “*the square is the head*”), so he/she has to give up this instruction and express it in another way (“*the square to the top*”), then it is accepted. In the second sequence, the user even tries to precisely define a certain amount like “*2cm is as long as a leg of the triangle*”.

1. The square is the head ... the square is the head of the figure ... the big triangle to the right ... stop ... upwards ... stop ... the square to the top ... stop ...

 2. The second big triangle rotates by 45 degrees to the left ... move to the right ... stop ... 2 cm to the top ... 2 cm is as long as a leg of the triangle ...

Fig. 2. Two sequences of subjects' commands

The development of the grammar was based on the Grammatical Framework (GF) [5], which is a grammar formalism for developing multilingual grammar applications.

Section 2 gives an introduction to the fragment analysis, including the utterance classification and the syntax analysis. Section 3 explains the formulation of the grammar, which includes the implementation of multilingualism and the composition of the grammar. Section 4 presents experimental results with corresponding discussion from two tasks: parsing and linearization. Section 5 concludes the paper.

2 Utterance Analysis

To get a better view of the utterances that appeared in the corpus, we decided to first classify the utterances in order to find out the common patterns and concepts used by

the subjects, and then through syntax analysis gather helpful information to find a good way to design the grammar accordingly.

From the 10 available transcripts of the NIMITEK corpus, 15 samples (one transcript may have one or two Tangram tasks) of Tangram were collected, which contain 2494 utterances. The Utterances were already manually annotated with three labels: command, comment, and question. In this work only ‘command’ utterances were considered, because more than 80% (2070) of the utterances in the collected samples are labeled with ‘command’.

2.1 Utterance Classification

The process of classifying utterances helped us to detect common and non-common language patterns that were used by subjects. The utterances were therefore classified into three groups: simple utterances, complex utterances, and others.

1. Simple utterances

Utterances are simple if they have a simple syntactic structure, particularly if they are short and elliptic, and their referenced concepts of the task domain can be identified directly. For example, “move the triangle to the left” has a simple syntax: a verb phrase built up from a noun phrase and a prepositional phrase. The referenced concepts have no modifications (e.g. size and position for object, speed for movement), and thus can be identified directly: an action “move”, an object “triangle”, and a direction “to the left”.

2. Complex utterances

Complex utterances also have a simple syntactic structure, but they demand more interpretation effort to identify the referenced concepts. This is usually due to the usage of modifications. For example, the utterance “the big triangle on the left” has also a simple syntax: a noun phrase built up from an adjective, a noun, and a prepositional phrase. However, there are two modifications of the noun “triangle”: the adjective that specifies a size “big”, and the prepositional phrase that specifies a position “on the left”. We have to consider these two modifications to identify the target object: the triangle, which is big and is on the left.

3. Others

Utterances that cannot be classified to either of the two types above are collected in this group. They have one or more of the following properties:

- too complex syntactic structures. For example, “Jetzt nimm bitte das Dreieck, was von den beiden das größte ist und markier das erstmal (now please take the triangle, which is the largest of the both and label it first)”.
- too complex content or rarely used concept. For example, “das Quadrat ist der Kopf (the square is the head)”. The concept “Kopf (head)” appears only once in the corpus, apparently, it is uncommon in the task.
- too little or no information to be used to get any useful interpretation. For example, “Ich befinde mich mit meinen Anweisungen wieder auf der linken Abbildung (I want to locate my position in the left side of the figure again)”. This utterance was

manually labeled as ‘command’ in the corpus. However, for the system it is not a real command. This is because it contains no explicit knowledge about the task, which is useful for the system to give its next action.

Through this classification we got a basic idea of the construction of the commonly used utterances. The semantic structure of most commands can be captured by a rule like: *object + direction + action*. Additionally, a word list for Tangram, which contains all words used for each semantic category and modifications was also collected from the corpus.

2.2 Syntax Analysis

Part-of-speech (POS) tagging was done to help us understanding the syntax of sentences. Context free grammar (CFG) analysis was done to prepare the extraction and construction of rules. Utterances selected from the NIMITEK corpus were tagged automatically by a regular expression tagger from the natural language toolkit (NLTK) [6] for German. For example:

- Das/DT oberste/JJ Dreieck/NN
(the top triangle)
- Bewegen/VB nach/IN unten/RB
(move down)

Here, *DT* means determiner, *JJ* means adjective or numeral, ordinal, *NN* means noun with singular form, *VB* means verb with base form, *IN* means preposition or conjunction, *RB* means adverb.

Through a bottom-up induction method, the corresponding context free grammar (CFG) [7] was built manually from POS tags to constituents. For example, for the two sentences above, the related CFG derivations are as follows:

- DT JJ NN → NP → S
- VB IN RB → VB PP → VP → S

From this experimental approach, we generated a basic command structure with three basic parts for designing a grammar for the typical Tangram commands from the NIMITEK corpus: *object*, *direction*, and *action*. In addition, the *object* could have modifications such as *size*, *position*, and *color*. The modification *position* has two possible types: object related position and global position. For example, “the triangle near the square” represents an object related position; “the left triangle” represents a global position. Object related position is used rather uncommon by subjects in contrast with global position, so we decided not to take the object related position into account for our grammar. Actions like movement could also have modification such as *speed*. For example, “go slowly”, “move fast”.

3 Grammar Formulation

Based on the collected syntactic structures, i.e. the rules for the naturally occurred utterances that were summarized from the corpus, we formulated the grammar.

3.1 Multilingualism

Besides the requirement of parsing commonly used user inputs, another major requirement of our grammar is multilingualism. In particular, we want to use one task-specific grammar for multiple languages. This is a specific problem of grammar engineering and can practically be solved by separating language dependent and language independent grammar components. The language independent part is called abstract syntax. It comprises concepts and rules of the task domain. A concrete syntax on the other hand, combines information about how to build up syntactical constituents from words for one specific language.

We chose GF [8] to implement our grammar, because it has many advantages for multilingual grammar applications. GF brings the Resource Grammar Library (RGL) [9] with it, which mainly provides language dependent and independent functions to form syntactical constituents. A wide range of languages is supported, including English and German. By using these functions, morphological aspects (e.g. inflection, gender and number agreement, long distance dependency...) are handled automatically, which eases the process of grammar engineering.

Additionally, GF uses a module system, which allows different languages to share declarations for common concepts (like noun-phrases) and building functions for them, while each uses its own definitions. This allowed us to use generic building functions to formulate the concrete syntax and outsource the dictionary of words (including knowledge about features like gender etc.). A concrete syntax is then called incomplete concrete syntax. It ensures two things, which also ease grammar engineering: first, changes of the concrete syntax will automatically affect all languages; second, adding new languages is achieved by simply adding an appropriate word lexicon.

To illustrate it more clearly we give examples for the particular syntaxes for Tangram:

1. Abstract syntax: The abstract syntax gives a set of declarations and functions for categories and production rules. Fig. 4 shows a fragment of the abstract Tangram syntax. **cat** means categories, **fun** means functions, i.e. grammar rules. The syntax defines four categories and one rule to form a command. The fragment in **fun** formalizes a rule that a command is made of an action (e.g. “flip”, “select”), an object (e.g. “the small triangle”), and a direction (e.g. “to the left”).

```

cat
Command;
Object;
Action;
Direction;

fun
CmdPhrase: Object --> Action --> Direction --> Command

```

Fig. 4. Example: abstract syntax for Tangram

2. Concrete syntax: Concrete syntax maps abstract syntax trees into linguistic objects. The objects can be simply strings or records that contain more strings, agreement features, etc. [10]. For instance, a part of the concrete syntax for Tangram is shown in Fig.5. **lincat** defines the linearization types, and **lin** defines the combination

rules for linearization, which state e.g. that *Command* should be an imperative verb phrase or that the object is a noun phrase. These categories are combined by using type conform building functions (e.g. `mkVP` ...) to form a command.

```

lincat
Command = Imp ;
Object = NP ;
Action = V ;
Direction = Adv ;

lin
CmdPhrase o a d = mkImp (mkVP (mkVP a o) d) ;

```

Fig. 5. Example: concrete syntax for Tangram

Furthermore, different languages have different parameter systems and lexicons. This ensures the independence of the grammar and the expansibility of the language.

3.2 Grammar Formalism

Following Ranta [11], the grammar in our work is composed by the following parts:

1. Abstract syntax: declare categories used by the user and functions in Tangram, i.e. semantic grammar rules.
2. Incomplete concrete syntax: define the declared categories and functions by using the language independent syntax building functions of the RGL, and uses a concrete syntax in 3.
3. Concrete syntax: declare the language dependent resources, which includes separate language syntax and word lexicon for German and English.
4. Lexicon interface: declare words, i.e. word symbols, and their types. The interface is language independent.
5. Lexicon instances: define words together with their morphological features (e.g. case, gender) for both German and English.

To get a better understanding of the parsing process, let us see a parsing example taken from the command line interface of GF shown in Fig. 6.

```

>parse "bewege das kleine Dreieck nach links" ["move the small triangle
to the left"]

Command PositivePol (Translate Simple (mkObjectNP PieceK
DefiniteSg (mkObjectCN PieceK (Size Small) Triangle)) (DirectionS
LeftAdv))

```

Fig. 6. Example of parsing

Fig. 7 shows a simplified syntax tree, which illustrates the parsing result above and gives a very simple and clear interpretation. Every node in the tree represents a syntactic constituent and is formed by the constituents of its child nodes. One such

building step is represented by a specific rule in the grammar. For example, **Move** represents a verb phrase, which is built of a noun phrase **Object** and a prepositional phrase **Direction**, with “move” as its central verb.

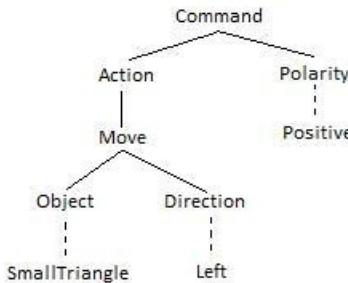


Fig. 7. Example of the syntax tree

4 Results and Discussion

We used 2070 utterances (cf. above) in the samples of Tangram in the NIMITEK corpus to test our grammar. The evaluation includes parsing, i.e. language analysis, and linearization, i.e. language generation.

4.1 Parsing

Since the grammar is currently not sensitive with respect to punctuation, we get the parsing results (shown in Table 1) by omitting all punctuations in the utterances.

Table 1. Results of utterance parsing

Command utterance	Number
Parsable	1248
Non-parsable	822

Parsable means everything that is acceptable by the grammar and thus by the parser. Utterances that are parsable include the two groups that we have mentioned earlier in utterance classification: simple utterances and complex utterances.

Non-parsable means everything that is not acceptable by the grammar. This includes utterances of the third group of the utterance classification that was mentioned in section 2.1. We decided not to build rules for them because they are hard to process (according to the three properties in this group) and bring little gain in our scenario. Otherwise, we found non-parsable utterances in the corpus, which could be turned into parsable utterances with little effort. For example, the parser fails to build a syntax tree, when there is an unknown word, non-agreeing word endings or wrong spellings. This could be repaired easily in most cases. However, we did not test this approach so far.

4.2 Linearization

Linearization means language generation, which generates language strings from syntax trees (according to the abstract syntax). It is the reverse process of parsing. A syntax tree can be rendered in any of the available languages (in our work: German and English) through concrete syntax. Fig. 8 shows a result of language generation for both English and German.

```
>linearize (Command PositivePol (Select (mkObjectNP PieceK
DefiniteSg Triangle)))
```

```
take the triangle
nimm das Dreieck
```

Fig. 8. Example of generation

Another important usage of linearization is translation. Different languages can be translated to each other. This is achieved by using a parse-linearize pipe: the input language will first be parsed with its grammar and lexicon, then linearized back to the target language with the target language grammar and its related lexicon. Fig. 9 shows a result of language translation from German to English.

```
>parse -lang=TangramGer "bewege das kleine Dreieck nach
links" | linearize -lang=TangramEng
```

```
move the small triangle to the left
```

Fig. 9. Example of translation

Here *TangramGer* is the German concrete syntax, and *TangramEng* is the English concrete syntax.

However, in our application, the advantage of the grammar is to support multilingual speech input and output, so we use it mainly for utterance parsing and generation.

5 Conclusion

In this work we developed a corpus-based multilingual grammar with the ability of both parsing and linearization. The grammar used self-defined semantic categories (e.g. object, action, direction) instead of syntactic categories (e.g. NP, VP). It was expected that the grammar can parse as many as possible naturally occurring utterances with simple syntax when solving Tangram puzzles interactively, and the goal was achieved by applying the following approach: First, classifying utterances to find out commonly used patterns and generating a word list of the task. Second, analyzing syntactic structures by applying POS tagging and CFG analysis. Third, formulating the grammar in GF, and synchronizing the syntax and lexicon in both German and English for multilingualism. The parser was integrated now in a Tangram demonstrator of the NIMITEK project with both speech and textual input interface.

The development was based on the transcripts of the NIMITEK corpus and the interface of the grammatical framework (GF). The transcripts of the NIMITEK corpus are in German, English transcripts are not available.

Additionally, during the analysis of the utterances, we found out that seven in ten subjects in the recordings of the NIMITEK corpus used simple, short instructions, while the other three used more complex, clause-based instructions. This may show a fact that most of the people intuitively choose to use simple or elliptical instructions, which on one side reduces the complexity of grammar development, but on the other side increases the degree of ambiguity, i.e. what users actually want or want to do. For instance, there are five triangles in Tangram, for user commands like “*the triangle*”, which one did the user really mean? Therefore, a corresponding dialogue strategy was planned and used for a backup. The strategy applied different types of questions for the system to guess and confirm the intention of users. For instance, giving a command like “*the triangle to the right*”, system will first guess the most possible triangle according to the state history of the task, then mark it and ask for user’s confirmation with a question like “*do you mean this one?*” In this way the system’s ability of solving ambiguity in natural language understanding was improved. Furthermore, if the user commands from the results of automatic speech recognition (ASR) are not understandable, according to the dialogue strategy the system outputs will be like “*pardon?*”, “*sorry, I do not understand you.*”, or “*I did not get it, could you please repeat it again?*”.

Acknowledgment. The presented study is based on work within the NIMITEK project (<http://wdok.cs.uni-magdeburg.de/nimitek>), and within the SFB TRR 62 ('Companion Technology for Cognitive Technical Systems') funded by the German Research Foundation (DFG). The responsibility for the content of this paper lies with the authors.

References

1. Gnjatovic, M., Rösner, D.: Inducing Genuine Emotions in Simulated Speech-Based Human-Machine Interaction: The NIMITEK Corpus. *IEEE Transactions on Affective Computing*, 132–144 (2010)
2. Fraser, N., Gilbert, G.N.: Simulating speech systems. *Computer Speech and Language* 5, 81–99 (1991)
3. Gnjatovic, M., Rösner, D.: Gathering Corpora of Affected Speech in Human-Machine Interaction: Refinement of the Wizard-of-Oz Technique. In: *Proceedings of the International Symposium on Linguistic Patterns in Spontaneous Speech (LPSS 2006)* (2006)
4. Gnjatovic, M., Rösner, D.: The NIMITEK Corpus of Affected Behavior in Human-Machine Interaction. In: *Processing of the Second International Workshop on Corpora for Research on Emotion and Affect* (satellite of LREC 2008), pp. 5–8 (2008)
5. Ranta, A.: Grammatical Framework: A Multilingual Grammar Formalism. *Language and Linguistics Compass* 3 (2009)
6. <http://www.nltk.org>
7. Knuth, D.E.: Semantics of Context-Free Languages. *Theory of Computing Systems* 2(2), 127–145 (1968)

8. Ranta, A., Angelov, K., Bringert, B.: Grammar Development in GF. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session (EACL 2009), pp. 57–60 (2009)
9. Ranta, A.: The GF resource grammar library. Linguistic Issues in Language Technology 2 (2009)
10. Khegai, J., Nordström, B., Ranta, A.: Multilingual syntax editing in GF. In: Gelbukh, A. (ed.) CICLing 2003. LNCS, vol. 2588, pp. 453–464. Springer, Heidelberg (2003)
11. <http://www.grammaticalframework.org/doc/gf-tutorial.html>

Part IV

Novel Interaction Techniques and Devices

Dancing Skin: An Interactive Device for Motion

Sheng-Han Chen¹, Teng-Wen Chang¹, and Sheng-Cheng Shih²

¹ Graduate School of Computational Design, National Yunlin University of Science and Technology, 123 University Road, Section 3, Douliou, Yunlin 64002, Taiwan, R.O.C.

² Department of Digital Design, Mingdao University, Taiwan, R.O.C.

{g9834710, tengwen}@yuntech.edu.tw, scshih@mdu.edu.tw

Abstract. Dynamic skin with its complex and dynamic characteristics provides valuable interaction device for different context. The main cause is the motion design and its corresponded structure/material. Starting with an understanding of skin/thus dynamic skin, we move to motion samples for case studies for unleashing the design process of motion in dynamic skin. The problem is to find a pattern of motion in dynamic skin. How to penetrate architectonic to cause the cortex to produce motion and we penetrates various types of street dance movement for motion design. This systemic skin construction can be a reference for building basic structure of folding form type skin and joint, developing motion which it needs, also provides dancer an interface that can interchange with other far-ended dancer through the Internet, regarding as a new manifestation and perform way for the street dance and its dancers.

Keywords: Dynamic skin, Motion, Folding form type, Street dance.

1 Introduction

Dynamic skin with its complex characteristics provides valuable interaction device for diverse context. However, most of dynamic skin designs, while pioneering the movable and interaction design are making specific cases either in interaction or architectural design. The main cause is the motion and its corresponded structure with material.

1.1 Dynamic Skin: From an Architecture Component to a Movable Interface

Originally, skin in architecture is a partition or façade that will reflect and protect habitants and their activities behind the skin. With movable technology such as robotic researches, a physical object has an ability to move and change its own shape according the material and surrounding contextual information. This creates a possibility for allowing the skin not only be partition but also a movable object [1]. Thus, a new type of skin with motion is developed called *dynamic skin*.

The maturity of sensor technology made that the surroundings can sense what people act that create an interacting behavior between people and their environment. It improves people's daily live and made it more convenient. With sensor technology, the dynamic skin has the capability to not just move but also react to the habitant around it[2]. This creates a new interaction within space.

1.2 Dynamic Skin Design Process

The concept of dynamic skin is simple but its design process is tedious and complex. Firstly the motion of a desirable dynamic skin needs to be defined and sometime preset. Then we use specific forms of dynamical skin just to show different design and its corresponded motion. Basically, there are two types of form of dynamical skin: unit cell organization and folding form, which will be introduced in some details in next section. After the type of form is determined, structure design then begins. Consequently, we had to test material and select specific Actuator and supported information to skin to create motion of skin. An ideal interactive skin should be combined with units as follows: 1) action joints and structure units, 2) sensor, and 3) actuator.

1.3 Motion in Dynamic Skin Design

With the characteristics of dynamic skin, motion is the essential part of dynamic skin design. Most of cases we found design their special motion based on the objectives of skins. According to the two types of skin forms, the design of structure of unit cell organization and folding form were very different. The motions of unit cell organization were more simply, they were not making influence from each other, and copy large number units to be a skin. The motions of folding form should design the structure connecting to each other, these structures influenced each other by action, and they could show variety of motion by just one structure. However, how to make a folding form structure that has variety motion the motivation of this research. In addition, finding a good example of motion will provide enough information for us to understand the motion of dynamic skin.

1.4 Street Dance as a Motion Paradigm for Exploring the Dynamic Skin Design

For finding possible motion example, street dance with its strong and direct movement is selected as our example of study. Street Dance is one of the most important dancing activities of teenager; it originated from Brooklyn in USA early 90s. The residents danced free on street, show their attitude of live by dancing. Dancers show every kind of dance poses by extremities and show the wave and variety motion of street dance. The motions of extremities of dancer were cooperated with every joint of human body, just like folding form skin that should use structure to make skin motion and do different motion again.

2 The Problem and How We Approach

From motion, we need to design out the appropriate construction. Most of cases are specific tailor-made structure with corresponded motion design. The problem of this research is to find a pattern of motion in dynamic skin design. Will there be patterns and how these motions affect the design outcome of dynamic skin and what are the design process focusing on the motion movement of dynamic skin?

Based on folding form type, a system called *Dancing Skin* is implemented with six steps: (1) interviewing the street dance dancer in advance, (2) analyzing dancer's

extremities and basic street dance movement, (3) designing folding form style of skin as the cortex construction, (4) giving correspondence between cortex's joint and dancer's extremities, and (5) formulating rules on corresponding also (6) putting this system in action, using Computer vision, Hear beat sensor, Distance Measuring sensor to detects dancer's movement, heartbeat, body position, and use it as input data for the system.

Dancing Skin provides the dancer to interflow with the wall by oneself, and obtains dance inspiration. It can also make interchange with the far-end dancer through the Internet, to produce more dance inspirations and cortex's motion.

3 Analysis

In folding form type, how to penetrate architectonic to cause the cortex to produce motion, even the whole structure to change, not only just simplex motion. In order to understand how to design the skin of folding form mode, let the skin have multiple motions, and penetrates various types of street dance movement as the research for motion design. To reach the goal above, we carry on to the following analyses and correspondences.

3.1 Motion in Dynamic Skin Structure Design

In this part we will discuss references of motion in folding form skin mode also look for the important factor in motion. The Fig. 1 (a)(b) is the Hylozoic Soil [3] designed by Philip Beesley in 2007, the architectonic was reticulate interconnected transparent join as Fig. 1(a), using the electric capacity sensor and the shape memory alloy driver, and causing the structure to perform periodically like feather as Fig. 1(b). The Fig. 1 (c)(d) [4] is the Beach animal by Theo Jansen, using lumber framework also making massive joints, start motion by wind power to let the installments walk on the beach as Fig 1(d). From the cases above, we found that the ideal structural design of folding form skin utilizes joints as link; the interaction of joints can produce chain-reacted motion.

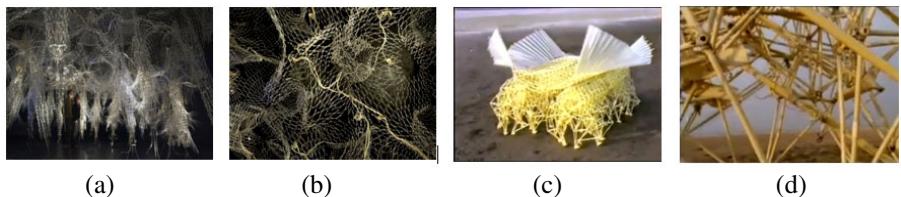


Fig. 1. The Hylozoic Soil (a), (b). The Beach animals (c), (d).

3.2 Street Dance Motion Gestures and Corresponded Motion Feedbacks

The analysis on the gestures of street dance motion is conducted via interviewing seven dancers with 2-7 years of dancing experience. The interview is conducted in face-to-face for collecting data in the later analysis. The gestures of each dancer are also recorded and interviews are based on exploring the gesture movement.

Based on the data collected, we divided the dancer's body into four parts of discussions and analysis: (a) Chest: The basic body part utilization for street dancer, chest movement makes the body to have height fluctuation, for dancer to display the rhythm, chest movement can be break down into four main movements: up, down, circle left, circle right. (b) Bottom: Link of main body part, upper and lower limbs, many movements must match with bottom in order to be smoother, these movements can be divided into: forward, backward, left, and right. (c) Hand: Hand movements increase visions of the dance, which divided into: Stretch out, retract, and wave. (d) Foot: The most nimble body part in dancing, the footstep movement can increase bounce and sprite in street dancing that divides into: slides, lifts, kicks, treads, cross, opens and closes five parts.

3.3 Chain-Reacted Motion with Interactive Behaviors

In the mode of folding form skin, motion is produced by continual influence, like the transmission of wave, starting from the front and passing on the back end, the reaction between dancer and skin shown in Fig. 2, so when the dancer is dancing, one's body uses the same way to achieve movements. The analysis of interactive behaviors is collected based on two conditions: single dance and battle-of-dance.



Fig. 2. The dancer and dancing skin

The consequential motion is analyzed from the interviews of above in addition to the information available on the web for the group or battle-of-dance. For example, a series of reaction on a single dancer is based on the gesture or action that dancer acts, such that the dancer made a wave movement starting from chest pass down to the bottom; the skin may also achieve such motion through the joints and reach the same kind of motion. The system is designed base on interaction between using dancer's continuous body movement for the skin to produce the same motion.

4 Implementing Dancing Skin

Dancing Skin is the prototype implemented for testifying the working concept and gathering the process information desired for studies. We used human joins and street dance extremity to design a skin that can dance, made skin approach variety motion mode. The Fig. 3 is the system interface diagram shows the concept of interaction

between dancer and skin. We linked wall and wall through Peer-to-Peer, so that dancers at different places can communicate with others via skin and provide more motion onto the skin.

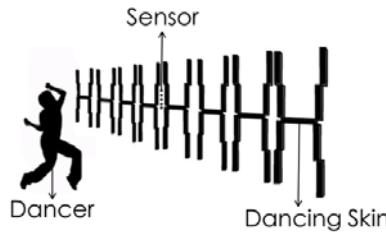


Fig. 3. Dancing Skin interface diagram

4.1 System Design

Dancing Skin is comprised of (a) Input: Gathering data detected by optical sensor, heartbeat sensor, Distance Measuring sensor for classification and rule base; (b) Operation processing: After processing sensors classification, and comparing with functions, produces motion information from skin in advance; (c) Output: After obtaining the information, using the skin controller to pass on to the actuation motor, enables the cortex to achieve motion. (Fig. 3)

When dancer starts dancing, the input sensors to detect the dancer's body, dividing into optical sensors, heartbeat sensor, distance measuring sensor1, distance sensor2, distance sensor3, the optical sensor will take the detected images using Classification to the processing unit finding the corresponding movement information, and start comparing to the numeric detected by other sensor. When other sensor detected information from second heartbeat sensor, the dancer's palpitation will affect the motor in area A of the skin, the louder the heartbeat, the bigger angle the motor transfer, and more obvious cortex motion to become. After receiving data from the heartbeat sensor, it will continue to match the next faction that is, using distance measuring sensor1to detect dancer's hand position, however, when sensor can not detected any hand information, it will skip to the next faction until all the factions have done matching and pass on the information above to the skin controller in output unit, then separate into different act, driving each part of the motor according to different sensor value, changing the motor's running angle, for skin to start motion.

4.2 Components

The skin is on the basic of H shapes and constructed by eight H (Fig. 4). The H shapes construction is divided into upper and lower limbs, corresponding to the four body parts with joints: hand, chest, bottom and foot, to simulate the motion directions. The joints are single direction joint and double direction joint. Single direction joint is situated at the top and the bottom end of the structure that can move back and forth. The double direction joint situated in the middle of the structure, which connected the

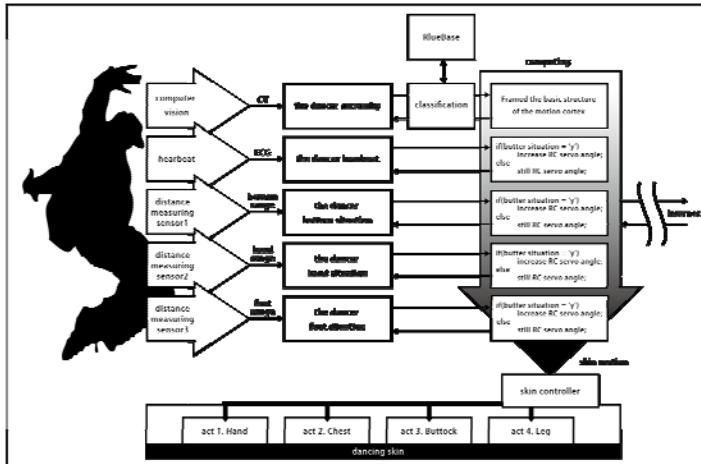


Fig. 4. The system diagram of dancing skin

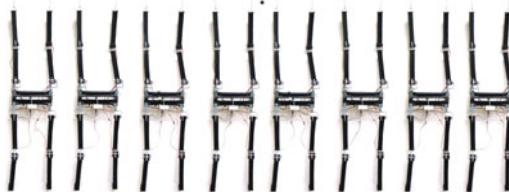


Fig. 5. The dancing skin entity

top to bottom with four junctions, this joint have the ability to move from front to back, right to left two directions, movement through these two kind of joints, causes the skin to has motion as Fig. 5.

The main system brace can be put into movement of front and back and movement of right and left that can be achieved by using RC servo, combining with the joints that will actuate the skin to make changes. Choosing the suitable material also plays an important part when designing the skin. After testing the material, we use cardboard for the material of the system brace. After determine using cardboard, intersecting two cardboards as a square for connecting with the motor joint to complete the construction of skin.

4.3 Interaction

Dancing skin sensed dancers bodies, and then produced skin motion as feedback. Dancers also could communicate with other dancers with network. The dancerA danced and sensed by SkinA that sent data to SkinB for data comparison, the RC Servo at SkinB got data and made SkinB produce motion, made dancerB get the feedbacks from dancerA. Dancers at two place produced interaction and more Skin or dancer motions are used as a reference for interaction.

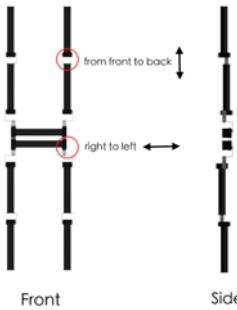


Fig. 6. The dancing skin joint move two directions

5 Two Scenarios

5.1 Dancing with Skin

Jadon was a busy office worker; he usually went off work at ten a clock in the evening. To a man who was a member of dance club at high school, Jadon still loved dance just like past, but he could not find a partner dance together after work. Jadon turned on the radio and danced to the skin. The skin started to sense the action of Jadon, then danced with Jadon. When Jadon did an open-close movement of hands and the feet, the skin body will also do a out turn so as the hand part, and then turn toward inside, as for the lower limb part's motion, the first joint of the leg part on skin will make a out turn and leading the second joint to move, too, then all turn inside again. However, when Jadon starts to make a cross step forward with his feet, the first and second of the lower limb joint on skin start to make a substantially cross step, corresponding dancer's open-close movement also display relation between skin and dancer's motion.

After a while, music which Jadon was practice at high school played. Jadon was familiar to the music and started to increase the strength of his dance step, his heartbeat became faster because of his emotion and dance step. The data all above that were sensed by sensor and made the angle of RC Servo increase, made the extent of skin motion increase. The skin motion became more and more clear until Jadon felt tired and his action became weak gradually. Jadon not bored to practice dance by himself any more, he could dance happily with skin.

5.2 Battle-of-Dance via Dancing Skins

Randy and Kevin was partner in dance club at high school, they always danced together. However they were hard to practice dance together since they entered college. With Dancing Skin, Randy and Kevin connect each other. Firstly, Kevin danced first, when Kevin danced was rhythmic, the extent of skin was large. When Randy saw the skin motion, he felt withdrawal because he was afraid that he could not dance as good as Kevin. So in Randy's turn, the extent of skin became smaller. Randy starts to think movement Kevin just did. Kevin did a body wave causing the skin to present a very attractive scope, so he also wants to give it a try. Randy made a

big motion from his chest down to his bottom, the first joint in the upper limb of the skin starts to move forward and finish at the last joint of lower limb, presenting a motion of wave. Randy's movement changes from small to big not only cause the response of the skin become more obvious, but also Kevin felt the message that Randy wish to send through the skin. Randy and Kevin did the communication of dance, they practiced dance together and shard the dance step of each other.

6 Conclusion

In this research, we combine folding form type and the street dance movement to design skin. Using human structure as reference, we discovered the relation between skin and dancer's body movement. The variation of street dancer's body can provide the skin to have diversity forms; this is also a reference for skin's motion design and knowing gearing relation of the skin's joints. We can see that the transformation between motions may penetrate through small movement, causes the next motion of the skin to be more obvious. How to design linking movement is another lesson learned, Further, This Dancing Skin construction provides a lesson for building folding form and joint, developing motion needed, also provides dancer an interface as a new manifestation and perform way for the street dance and its dancers.

References

1. Sterk, T.: Shape Change in Responsive Architectural Structures: Current Reasons & Challenge (2006)
2. Chang, T., Lin, H.-H., Shih, J.-H.: An emotional Interface for Smart Living Space, Space Living-Living Space. Taiwan Society of Digital Media Design, Douliou, Yunlin, Taiwan (2008)
3. Beesley, P.: Hylozoic Soil Geotextile Installations 1995 - 2007. Riverside Architectural Press (2007)
4. Jansen, T.: The Great Pretender. 010 Uitgeverij (2007)

A Hybrid Brain-Computer Interface for Smart Home Control

Günter Edlinger, Clemens Holzner, and Christoph Guger

Guger Technologies OG and g.tec medical engineering GmbH
Herbersteinstrasse 60, A-8020 Graz, Austria
edlinger@gtec.at

Abstract. Brain-computer interfaces (BCI) provide a new communication channel between the human brain and a computer without using any muscle activities. Applications of BCI systems comprise communication, restoration of movements or environmental control. Within this study we propose a combined P300 and steady-state visually evoked potential (SSVEP) based BCI system for controlling finally a smart home environment. Firstly a P300 based BCI system was developed and tested in a virtual smart home environment implementation to work with a high accuracy and a high degree of freedom. Secondly, in order to initiate and stop the operation of the P300 BCI a SSVEP based toggle switch was implemented. Results indicate that a P300 based system is very well suitable for applications with several controllable devices and where a discrete control command is desired. A SSVEP based system is more suitable if a continuous control signal is needed and the number of commands is rather limited. The combination of a SSVEP based BCI as a toggle switch to initiate and stop the P300 selection yielded in all subjects very high reliability and accuracy.

Keywords: Brain-Computer Interface, Smart Home, P300, SSVEP, electroencephalogram.

1 Introduction

Human-Computer interfaces can use several different signals from the body in order to control external devices which can be based on muscle activity (EMG-Electromyogram), eye movements (EOG-Electrooculogram), respiration or heart rate variability. Recent improvements in terms of usability and reliability in normal subjects as well as handicapped persons allow now the usage of electrical brain activity (EEG-Electroencephalogram) as input signals. EEG-based brain-computer interface (BCI) systems have been realized on various phenomena of the awake brain: (i) slow cortical potential shifts [1], (ii) the P300 response [2;3], (iii) steady-state visually evoked potentials (SSVEP) [4;5] or (iv) somato-sensory rhythm (SMR) based i.e. motor imagery [4;6].

For the control of a smart home environment evoked potential BCI approaches are the most suitable ones because these approaches allow to select certain target commands out of many different commands to initiate a control. A further big advantage of these approaches is that the user can be trained within a short period of

time just with a very small sub-set of the possible selections. This means if the smart house has in total 200 control option, the BCI system can be trained on only 5 different icons. This allows the BCI system already to distinguish and classify between the 200 functions with high accuracy and relatively high speed (5-30 seconds per decision).

Another obstacle found in BCI literature is the fact that a certain percentage of the population cannot operate a specific type of BCI due to various reasons. Inter-subject as well as intra-subject variability often leads to a so-called BCI illiteracy [4]. Across the different BCI approaches around 20%-25% of subject are unable to control one type of BCI in a satisfactory way [3]. Therefore, the usage of 'hybrid' BCIs has been introduced into the literature to overcome these problems using the output of somatosensory rhythm BCI as well as P300 or steady state visually evoked potentials based BCIs enabling subjects to choose between these different approaches for optimal BCI control [7].

A study of Hong et al. recently did a comparison of an N200 and a P300 speller (tested on the same subjects) and found similar accuracy levels for both of them [8]. This gives evidence that a closer look to the N200 component could be promising, at least for some subjects. Hence BCI illiteracy could be overcome or maybe minimized by investigating more thoroughly subject specific preferences. The group of Kansaku reported about the improvement of BCI P300 operation using an appropriate color set for the flashing letters or icons [9]. However, previous studies using the P300 BCI approach for the control of devices within a virtual smart home environment [10] indicate that such an evoked response based BCI can be reliably utilized. Interface masks having different complexity depending upon the capability of the devices can be operated. Another issue in BCI control is to tackle the so called zero class problem [7;11]. A P300 speller for example can be operated successfully from a high percentage of the population with high accuracy and reliability. However, starting and stopping, i.e. switching on and off the BCI operation is still done manually. Both, P300 and SSVEP BCIs were selected for the study setup as recent studies indicate [12] that also severely handicapped people could operate a P300 BCI in a satisfactory way. Allison et al [13] and Zhang et al. [14] showed that only selective attention onto a pattern alone is sufficient for SSVEP based BCI control . The latter paper achieved an overall classification accuracy of $72.6 +/- 16.1\%$ after 3 training days. Therefore also severely disabled people, who are not able to move their eyes, can control an SSVEP-based BCI.

The current study introduces the usage of a hybrid BCI approach for optimizing control comfort of certain interface masks i.e. using (i) the P300 approach for when one selection out of many classes have to be done and (ii) using an SSVEP based toggle switch to start and stop the P300 BCI. As a test bed environment various domotic devices in a smart home environment were controlled.

2 Combined P300 and SSVEP BCI Approach

2.1 P300 Base System

The P300 spelling device is based on a rectangular matrix layout of different characters or icons displayed on a computer screen. A single character or icon is

flashed on and off in a random order as shown in Fig. 1A. The underlying phenomenon used to setup a P300 speller is the P300 component of the EEG, which is elicited if an attended and relatively uncommon event occurs. The subject must concentrate on a specific icon he/she wants to select. When the icon flashes on, a P300 component is induced and the maximum in the EEG amplitude is reached typically 300 ms after the flash onset. Such a P300 signal response is more pronounced in the single character speller than in the row/column speller and therefore easier to detect [3].

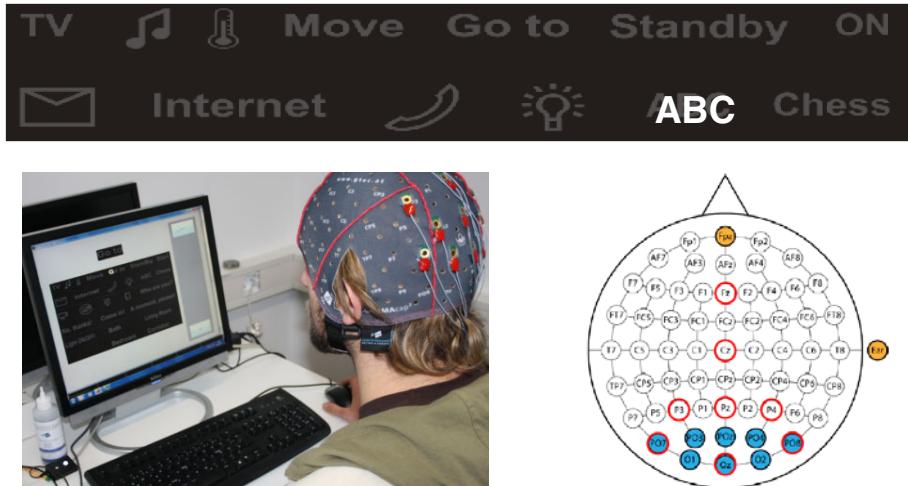


Fig. 1. The upper panel displays an example of the BCI user interface mask with the standby icon ('Standby') and keyboard icon for spelling ('ABC'). The lower left panel displays a subject equipped with an active electrode system (g.GAMMAAsys, g.tec medical engineering GmbH, Austria) operating the SSVEP-P300 interface mask. The lower right panel indicates the used electrode positions. A total of 13 electrode positions are mounted over the parietal and occipital areas according to the international extended 10/20 system.. The red color coded electrodes (Fpz, Cz, P3, Pz, P4, PO7, Oz, PO8) are utilized for P300 operation and the blue color coded positions (PO7, PO3, POz, PO4 PO8, O1, Oz, O2) for SSVEP operation. The ground electrode is positioned at Fpz and the right ear lobe is utilized for the reference electrode.

For BCI system training, EEG data are acquired from the subject while the subject focuses on the appearance of specific letters in the copy spelling mode. In this mode, an arbitrary sequence of icons is presented on the monitor. First, the subject counts whenever the first target icon flashes. Each icon is flashed on for about 100 ms per flash. Then the subject counts whenever the second target flashes until it flashes 15 times, and so on. EEG data are evaluated with respect to the flashing event within a specific interval length, processed and sent to a linear discriminant analyzer (LDA) to separate the target icons from all non targets. This yields a subject specific weight vector WV for the real-time experiments. It is very interesting for this approach that the LDA is trained only on e.g. 5 icons representing 5 classes and not on all possible classes in the mask (details about P300 speller setup can be found in [3]).

2.2 SSVEP Base System

SSVEP based BCI system use flickering lights (LEDs) or flickering symbols on a normal computer screen to visually stimulate the user with a certain flashing frequency between 5 up to 25 Hz. If a light source is flickering with e.g. 14 Hz and the user is looking at it, then an EEG signal with an increased power at the stimulation frequency will be evoked over the occipital areas and can be made visible in the power spectrum of the EEG data (see Fig. 2). The evoked signal power drops down if the stimulation frequency increases. For a fixed stimulation frequency a simple threshold criterion can be used to determine if the user is looking at the light source, otherwise a LDA can be trained with the individual data to find the optimal threshold. If the number of light sources is increased a multi-dimensional control can be realized.

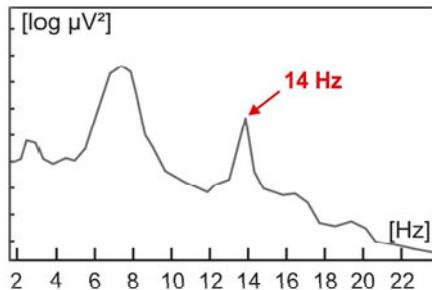


Fig. 2. Power spectrum of the EEG with a peak at the stimulation frequency at 14 Hz. The x axis displays the frequency in Hz and the y axis yields the log power.

2.3 Virtual Reality Smart Home Test Bed Setup

In order to operate the BCI control in the virtual environment several components have been developed. (i) biosignal amplifiers must be able to work in such a noisy environment; (ii) the recordings should ideally be done with a rather small portable device to avoid collisions and irritations within the environment; (iii) the BCI system must be coupled with the VR system for real-time experiments and (iv) a special BCI communication interface must be developed to have enough degrees of freedom available to control the VR system. Fig. 3 illustrates the necessary components in detail. A 3D projector is located next to a projection wall for back projections. The subject can be positioned in front of the projection wall to avoid shadows and is equipped with position tracker to capture movements, shutter glasses for 3D effects and the biosignal amplifier including electrodes for EEG recordings. The XVR (eXtreme VR, VRmedia, Pisa, Italy) PC is controlling the projector, the position tracker controller and the shutter glass controller. The biosignal amplifier is transmitting the EEG data to the SSVEP - P300 BCI system which is connected to the XVR PC via UDP connection to exchange control commands.

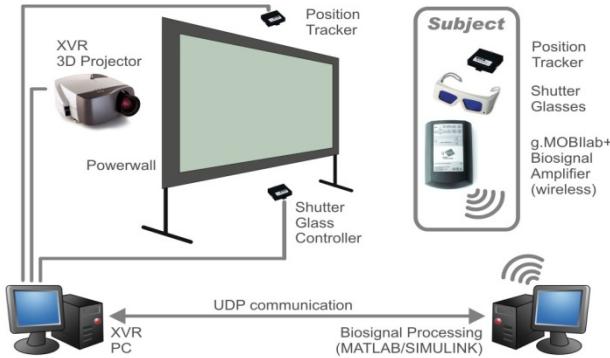


Fig. 3. Scheme of virtual environment setup

The virtual smart home itself consists of different rooms whereby each room is equipped with several different devices that can be controlled: TV, MP3 player, telephone, lights, doors, etc. Therefore, all the different commands were summarized in 7 control masks: a light mask, a music mask, a phone mask, a temperature mask, a TV mask, a move mask and a go to mask. Fig. 4 shows the TV mask and as an example the corresponding XVR image of the living room. The subject can e.g. switch on the TV by selecting the TV symbol. Then, the TV station and the volume can be regulated. For further details see [10]. In such an application, precise timing between the appearance of the symbol on the screen and the signal processing unit is very important. Therefore, the flashing sequence was implemented under Simulink where the real-time BCI processing was also running. Fig. 5 shows a Simulink model processing the EEG data in real-time and combining BCI control to the virtual smart home environment.



Fig. 4. Left panel: TV interface mask. Right panel: Example of the virtual living room displaying domotic devices to be operated like the TV set, music set, room light or chess board.

The signal and processing flow in Fig. 5 starts from the left hand side and progresses to the right hand side. The biosignal amplifier g.USBamp (g.tec, medical engineering GmbH, Austria) is reading 13 EEG channels into the model and is pacing the real-time application. The 'Source Derivation' block splits the channels and sends 8 EEG channels to the 'Signal Processing SmartHome' block for P300 control and

another 8 EEG channels to the 'SSVEP Processing' block. For the P300 processing chain data are band-pass filtered and downsampled to 64 Hz and the Signal Processing SmartHome block is performing the feature extraction and classification for the P300 system. The 'Control Flash SmartHome block' controls the icons representing the User Interface. The 'Sockets SmartHome' block send the specific commands to the smart home XVR control server via a UDP connection.

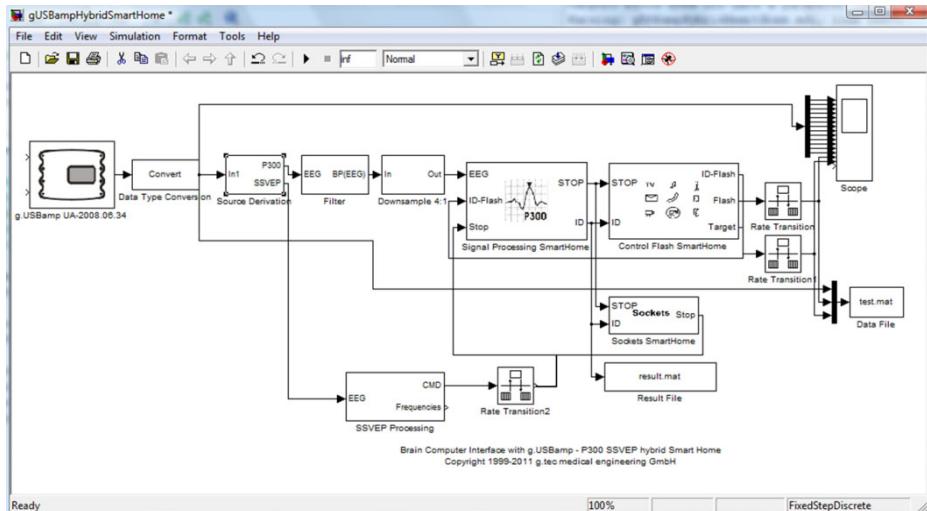


Fig. 5. Simulink real-time processing model of the combined P300 and SSVEP system

The 'SSVEP Processing' processes the EEG channels utilizing the Minimum Energy (ME) algorithm [15]. ME in principal projects artificial signal oscillations at the fundamental stimulation frequencies as well as the 1st and 2nd harmonics onto the orthogonal complement of the EEG-signals. The algorithm further combines the input channels in such a way, that the outcome energy is minimized. The presence of the stimulation stimulus and thus the attention of the user to the flashing light is determined by a test statistics which calculates the ratio between the signal with an estimated SSVEP-response and the signal where no visual stimulus is present. The output of the block is finally used to switch on and off the BCI system.

A total of 3 healthy subjects all right handed males between 25 and 36 years with no contraindication for observing flickering lights operated the combined P300 - SSVEP setup in the virtual smart home environment. To be able to measure the SSVEP signal electrodes must be mounted over parietal and occipital sites of the cortex as shown in Fig. 1 lower panel. The SSVEP method requires 8 EEG electrodes to show a high classification accuracy. The P300 uses 8 EEG electrodes over frontal, central, occipital and parietal sites and has 3 electrodes in common with the SSVEP principle. Therefore in total 13 electrodes will be investigated. The active electrode system g.GAMMASys was mounted according to the electrode position given in Fig. 1 lower panel and EEG data were sampled at 256 Hz using g.USBamp.

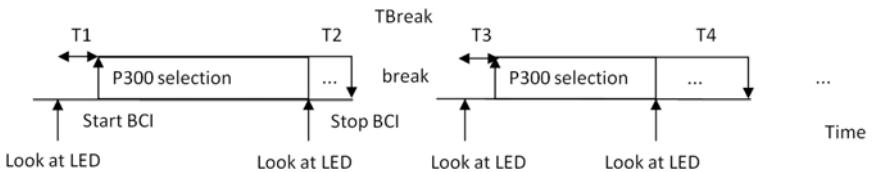


Fig. 6. Experimental paradigm and timing for the SSVEP - P300 experiment

A bright LED light source was connected to a special programmable stimulation device. The stimulation frequency was set via USB connection and a Simulink interface to 14 Hz. Subjects had to follow a certain paradigm in order to test the reliability and performance of the system (see Fig. 6). In order to start the operation of the BCI subjects were instructed to concentrate and look to the 14 Hz flashing light which is used here as a ON/OFF toggle button. After the 14Hz activity has been detected by the ME algorithm from the EEG (time T1), subjects had to select 5 predetermined commands from the P300 speller mask. Icons were flashed on in this case for 15 times so the P300 operation time has been fixed. Then the subjects had to operate the SSVEP toggle button to switch off the speller (switching off time T2). Before the next trial was started subjects had to wait for approximately one minute (TBreak) in order to determine false positive operations of the toggle button. An operator instructed the subjects verbally to continue the BCI operation by looking again at the flashing light to switch on the speller again. This procedure was repeated for 10 rounds of SSVEP - P300 operation.

3 Results and Discussion

Table 1 displays the results for the SSVEP controlled toggle switch and P300 accuracy for 3 subjects. T SSVEP on/off yields the mean time needed for switching on and off the BCI operation. All subjects needed about 4.5-5 seconds to switch on the BCI toggle button and about 5.8 to 7.9 seconds to switch it off again. Spelling accuracies were 100% for the 5 target icons. In the break in between the 10 spelling experiments only S3 displayed one false positive result.

BCI enabled control and communication is a new skill a subject has to learn. In an initial adaptation phase the BCI system is trained to the specific subjects brain activity. In addition the subjects have to get used and adapt to the BCI system as well. The time needed for a subject to adapt to the system is by far shorter in exogenous BCIs like P300 and SSVEP approaches [3;4]. Such BCI systems yield higher accuracies in a higher number of subjects and give therefore for control purposes more reliable results. The current study introduces a combined thus sequential usage of 2 types of BCI concepts.

Such an approach has the advantage that users might benefit from a more optimal performance of the overall system as subparts of the system are based on the most suitable control approaches. Pfurtscheller et al [7] introduced a SMR based BCI as brain switch. However, it is known from the literature that a reliable control of SMR

Table 1. Mean switch on/off time, P300 accuracies and number of correct operation of the toggle button

	T SSVEP On/Off [s]		P300 Accuracy [%]	Number of correct operation of toggle switch	
	On	Off		TP	FP
S1	4,6 Mean 6,2	7,8	100	10	0
S2	4,48 Mean 5,14	5,8	100	10	0
S3	5,05 Mean 4,68	7,92	100	11	1

activity needs a long training period [3]. Furthermore in a high percentage of the population such an approach could not be used. In contrast to SMR BCIs evoked potential based BCIs showed a better overall performance and reliability. However, focusing a very long time to flickering light sources, either flashing at distinct frequencies for SSVEP approach or flashing randomly for the P300 approach might distract people from their daily activities or simply annoy users after some time. Therefore the SSVEP approach might be utilized to operate a simple on/off toggle button in a very reliable way. Hence the BCI operation can be started and stopped arbitrarily by the user without the need of an operator intervention. In the current experiments only one false positive toggle switch event within the total of 30 min forced breaks was observed for SSVEP - P300 BCI. The operation of a speller like interface, i.e. the operation and selection of target commands out of many commands works in a very reliable way based on the P300 approach. In Guger et al. the spelling performance of a total of 100 subjects was investigated and more than 90% of the subjects could operate the P300 speller with 100% accuracy [3]. Edlinger et al. reported on the usage of the P300 BCI in the smart home environment [10]. There the authors concluded that the performance of the P300 control is comparable to the classical 6x6 speller. However, the authors also state that designing the interface masks in a more proper way can improve the usability and success rate in BCI. Moreover for simple control masks with less symbols to select like moving e.g. a device in one out of four directions the SSVEP based control might be more reliable and faster. Results of the current study suggest that a combined or hybrid BCI approach such as using the P300 BCI approach for a many class selection task and using the SSVEP especially as a toggle switch to initiate and stop BCI operation is promising. Furthermore the SSVEP interface can enhanced in a straightforward manner by adding other control flickering lights to improve the performance for e.g. a four class selection task within the smart home environment.

4 Outlook

Based on the experiments in the virtual smart home, the BCI system is currently further advanced to be used within a real smart home developed for independent living for handicapped people. Here the BCI system is embedded in a middleware platform that allows controlling multiple domotic devices with the BCI system.

Acknowledgments. Funded partly from the EC grant contract FP7/2007-2013 under BrainAble project, FP7-ICT-2009-247935 (Brain-Neural Computer Interaction for Evaluation and Testing of Physical Therapies in Stroke Rehabilitation of Gait Disorders) and contract IST-2006-27731 (PRESENCCIA).

References

1. Birbaumer, N., Ghanayim, N., Hinterberger, T., Iversen, I., Kotchoubey, B., Kubler, A., Perelmouter, J., Taub, E., Flor, H.: A spelling device for the paralysed. *Nature* 398, 297–298 (1999)
2. Sellers, E.W., Krusienski, D.J., McFarland, D.J., Vaughan, T.M., Wolpaw, J.R.: A P300 event-related potential brain-computer interface (BCI): the effects of matrix size and inter stimulus interval on performance. *Biol. Psychol.* 73, 242–252 (2006)
3. Guger, C., Daban, S., Sellers, E., Holzner, C., Krausz, G., Carabalona, R., Gramatica, F., Edlinger, G.: How many people are able to control a P300-based brain-computer interface (BCI)? *Neuroscience letters* 462, 94–98 (2009)
4. Allison, B., Luth, T., Valbuena, D., Teymourian, A., Volosyak, I., Graser, A.: BCI Demographics: How Many (and What Kinds of) People Can Use an SSVEP BCI? *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 18(2), 107–116 (2010)
5. Friman, O., Volosyak, I., Graser, A.: Multiple channel detection of steady-state visual evoked potentials for brain-computer interfaces. *IEEE Trans. Biomed. Eng.* 54, 742–750 (2007)
6. Pfurtscheller, G., Neuper, C., Muller, G.R., Obermaier, B., Krausz, G., Schlogl, A., Scherer, R., Graimann, B., Keinrath, C., Skliris, D., Wortz, M., Supp, G., Schrank, C.: Graz-BCI: state of the art and clinical applications. *IEEE Trans. Neural Syst. Rehabil. Eng.* 11(2), 177–180 (2003)
7. Pfurtscheller, G., Allison, B.Z., Brunner, C., Bauernfeind, G., Solis-Escalante, T., Scherer, R., Zander, T.O., Mueller-Putz, G., Neuper, C., Birbaumer, N.: The Hybrid BCI. *Front Neurosci.* 4, 42 (2010)
8. Hong, B., Guo, F., Liu, T., Gao, X., Gao, S.: N200-speller using motion-onset visual response. *Clin. Neurophysiol.* 120, 1658–1666 (2009)
9. Komatsu, T., Hata, N., Nakajima, Y., Kansaku, K.: A non-training EEG-based BMI system for environmental control. *Neurosci. Res.* 61(suppl.1), S251 (2008)
10. Edlinger, G., Holzner, C., Groenegress, C., Guger, C., Slater, M.: Goal-Oriented Control with Brain-Computer Interface. In: *HCI 2009*, vol. 16, pp. 732–740 (2009)
11. Haihong, Z., Cuntai, G., Chuanchu, W.: Asynchronous P300-Based Brain–Computer Interfaces: A Computational Approach With Statistical Models. *IEEE Transactions on Biomedical Engineering* 55(6), 1754–1763 (2008)

12. Nijboer, F., Sellers, E.W., Mellinger, J., Jordan, M.A., Matuz, T., Furdea, A., Halder, S., Mochty, U., Krusinski, D.J., Vaughan, T.M., Wolpaw, J.R., Birbaumer, N., Kubler, A.: A P300-based brain-computer interface for people with amyotrophic lateral sclerosis. *Clin. Neurophysiol.* 119, 1909–1916 (2008)
13. Allison, B.Z., McFarland, D., Schalk, G., Zheng, S.D., et al.: Towards an independent brain-computer interface using steady state visual evoked potentials. In: *Brain-computer interface systems: progress and prospects BCI Meeting 2005—workshop on signals and recording methods*, pp. 1388–2457 (2005)
14. Dan, Z., Xiaorong, G., Shangkai, G., Engel, A.K., Maye, A.: An independent brain-computer interface based on covert shifts of non-spatial visual attention 539–542 (September 3, 2009)
15. Friman, O., Volosyak, I., Graser, A.: Multiple Channel Detection of Steady-State Visual Evoked Potentials for Brain-Computer Interfaces. *IEEE Transactions on Biomedical Engineering* 54(4), 742–750 (2007)

Integrated Context-Aware and Cloud-Based Adaptive Home Screens for Android Phones

Tor-Morten Grønli¹, Jarle Hansen², and Gheorghita Ghinea¹

¹ The Norwegian School of Information Technology,

Schweigaardsgt. 14, 0185 Oslo, Norway, and

School of Information Systems, Computing and Mathematics, Brunel University,
Uxbridge UB8 3PH, London, United Kingdom

² School of Information Systems, Computing and Mathematics, Brunel University,
Uxbridge UB8 3PH, London, United Kingdom
george.ghinea@brunel.ac.uk, jarle.hansen@brunel.ac.uk,
tmg@nith.no

Abstract. The home screen in Android phones is a highly customizable user interface where the users can add and remove widgets and icons for launching applications. This customization is currently done on the mobile device itself and will only create static content. Our work takes the concept of Android home screen [3] one step further and adds flexibility to the user interface by making it context-aware and integrated with the cloud. Overall results indicated that the users have a strong positive bias towards the application and that the adaptation helped them to tailor the device to their needs by using the different context aware mechanisms.

Keywords: Android, cloud computing, user interface tailoring, context, context-aware, mobile, ubiquitous computing, HCI.

1 Introduction

The home screen in Android phones is a highly customizable user interface where the users can add and remove widgets and icons for launching applications. This customization is currently done on the mobile device itself and will only create static content. The content, buttons and widgets, are shown in the same position with the same icons until the user manually changes it. Our work takes the concept of Android home screen [3] one step further and incorporates this in regular Android applications. We add flexibility to the user interface by making it context-aware and integrated with the Google cloud. Cloud computing is focused on sharing data and computation resources over a scalable network of nodes. It has gained popularity over the last few years and large companies like Microsoft, Google and IBM all have initiatives promoting cloud computing [1]. The configuration of the application home screen is stored in the Google cloud and the server will push events to the registered mobile devices that are executed on the phone. Mark Weiser [2] argued that machines should fit the human environment and not force the humans to enter theirs. We try to follow

this idea by registering the context of the user from many different sources and automatically adapting the home screen based on the predefined user configuration.

The paper is organized as follows: Section 2 describes the research background, followed by application design and implementation in section 3. Section 4 addresses the preliminary results from our work and section 5 concludes the paper.

2 Background

Developers and researchers agree that context is an increasingly important factor when designing new mobile applications. Context-aware tailoring of information sources for end users can greatly increase the perceived use and agility of mobile environments.

Context-awareness in mobile applications has caught the attention of researchers [12, 13] and proves to be more and more an invaluable resource. For example, Ludford et al. [14] looked at the use of context to provide useful information to the user on their mobile phone. In their work, context is based on the location and/or time of the day. Another definition of context is the user's planned activity in combination with the location. This is interesting because it generates quite a lot of information about the user, but information is of reduced interest if we do not combine it with other contextual dimensions or aggregate the data. On an overall basis the use of context in applications is often missing or single dimensional. This focus should be changed, since automated information aggregation from contextual sources would possibly be able to not only support the everyday tasks of the user, but also improve efficiency and ease the work of the user by automatically tailoring information to the user's needs and/or adapting the application to the user's current setting. The widespread use of smart, mobile devices have led researchers and developers to consider context as an important criteria for highly mobile systems. The notion of context-aware computing is generally the ability for the devices to adapt their behavior to the surrounding environment, hence enhancing usability [15].

Towards this goal, Dey and Abowd [2] state that if we understand context fully in a given environment and setting, we would then be able to better choose what context-aware behaviors to sustain in applications. This could lead to more realistic applications and thereby applications more meaningful to users. Edwards [4] exemplifies this when he uses context information to build an application. In his application different layers represent different sources of information and they can be reused in later settings. Edwards argues that context is a major part of our daily life and that computing with support for sharing and using contextual information (context-aware computing) would improve user interaction. Indeed when viewing people rather than systems as consumers of information, a new infrastructure is needed.

The implementation of sensors is the second half of our context-aware dimension. By enabling such services we facilitates exploitation of new resources in a mobile environment. Traditionally sensors are an important source of information input in any real world context and several previous research contributions look into this. The work of Parviainen et al [7] approaches this area from a meeting room scenario. Here, we can find a variety of applications in which a sound source localization system may be useful, such as: Automatic translation to another language, retrieval of specific topics,

and summarization of meetings in a human-readable form. The paper describes briefly the source localization system developed for these tasks as well as evaluating the results from preliminary experiments. They find sensors a viable source of information, but also acknowledge there is still work to do, like improving integration.

Another contribution focusing on the information retrieval possibilities is the work of Chatzigiannakis et al. [8]. They study how wireless sensor and actor networks are comprised of a large number of small, fully autonomous computing, communication, sensing and actuation devices, with very restricted energy and computing capabilities. Further they investigate how such devices co-operate to accomplish a large sensing and acting tasks such as information gathering from real world events.

In line with our research Boulis et al [9] look into how sensor usefulness changes in response to the dynamic needs of multiple users. They propose a framework, SensorWare, to define and support lightweight and mobile control scripts that allow the computation, communication, and sensing resources at the sensor nodes to be efficiently harnessed in an application-specific fashion. Their proposal seeks to remedy the limited flexibility problem at the expense of increased responsibility for the programmer.

By taking the aspect of sensors and context-aware information and integrating it in a mobile application we acquire desirable effects for the end users. When this is further combined with remote configuration from cloud-deployed applications the benefits, could be even greater. Cloud computing focuses on sharing data and makes it possible to distribute storage and computations over a scalable network of nodes. Large IT companies like Microsoft, Google and IBM, all have initiatives relating to cloud computing [10]. One of the key features under this paradigm is scalability on demand, where the user pays for the amount of computation and storage that is actually used. Moreover, this scalability is usually completely transparent to the user. Mei et al. [10] have focused on finding interesting topics for future research in the area of cloud computing, and have drawn analogies between cloud computing and service and pervasive computing. They identify four main research areas: 1) *pluggable computing entities*, 2) *data access transparency*, 3) *adaptive behaviour of cloud applications* and 4) *automatic discovery of application quality*. Our work is closely related to *adaptive behaviour of cloud computing applications*. In our work we implement context-awareness in a mobile environment. Further we take advantage of remote configuration of in application settings, making the local application adept to changes saved in the cloud and pushed to individual attached devices.

One important issue with cloud computing is privacy. Data is often stored on external servers that one has no control over. Mowbray and Pearson [11] have looked at how it can be possible to implement a client-based privacy manager in a cloud-computing environment. Their solution reduces this risk by helping users to control their own sensitive information. This is implemented using an obfuscation and deobfuscation service to reduce the amount of sensitive information held within the cloud. They have also implemented a feature to allow users to express privacy preferences about the treatment of the personal information. Security is also an important issue in our work when dealing with sensitive user data. We address this by using state of the art Open Authentication services implementation in our application and integrating with the Google login process.

To summarize, our research investigates integrated context-aware and cloud-based adaptive application home screens for android phones.

3 Design and Implementation

The user starts by logging in to a webpage using his Google account, s/he is then presented with several options that will customize the home screen on his mobile device. One example is that the user can configure that the screen brightness on his mobile should be automatically lowered when the battery charge is below 20%.

Another example is that the user is able to select the applications he wants to be available on the home screen of his mobile device. By selecting for example the *Maps* application and pressing *Save configuration* this will be saved in the cloud and automatically pushed to his phone, figure 1.

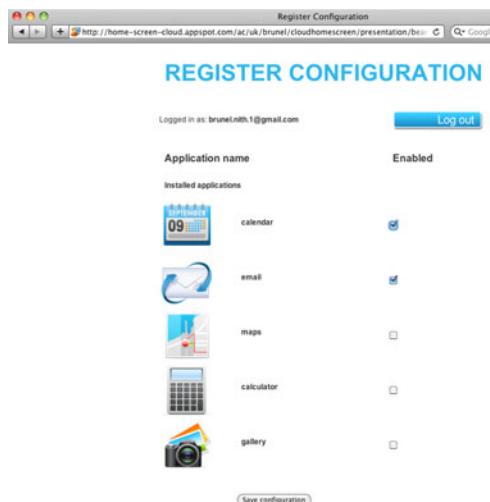


Fig. 1. Server overview

The phone will update the home screen when the push message is received from the Android Cloud to Device Messaging system. By configuring these options on the cloud we get several benefits. The configuration values are not tied to one specific device, so there is no need to manually configure each device. By registering the device in the cloud application, any predefined configuration values are pushed to the phone. It is also easier to add more advanced configuration options when the user can take advantage of a bigger screen, mouse and keyboard than those found on small or mobile devices.

The context-aware part of the system handles multiple input sources. The simplest form is the context taken from the phone, where battery level and screen brightness is two of the input options used in our application. We also integrated the system with Google Calendar. We added special tags, `${type=work}` or `${type=leisure}` , in the

description field of the scheduled meetings to describe the context. If the tag `${type=work}` was added, this lets the application know that the user is in a work setting and it will automatically adapt the contacts based on this input. In a work context only work related contacts would be shown.

The user interface in our solution consisted of two main components: 1) The home screen on the mobile device and 2) the webpage to add the configuration options. The home screen is of course the most important part. This screen handles most of the user interaction with the phone and all other applications can be launched from this screen. An overview of the system is shown in figure 2.

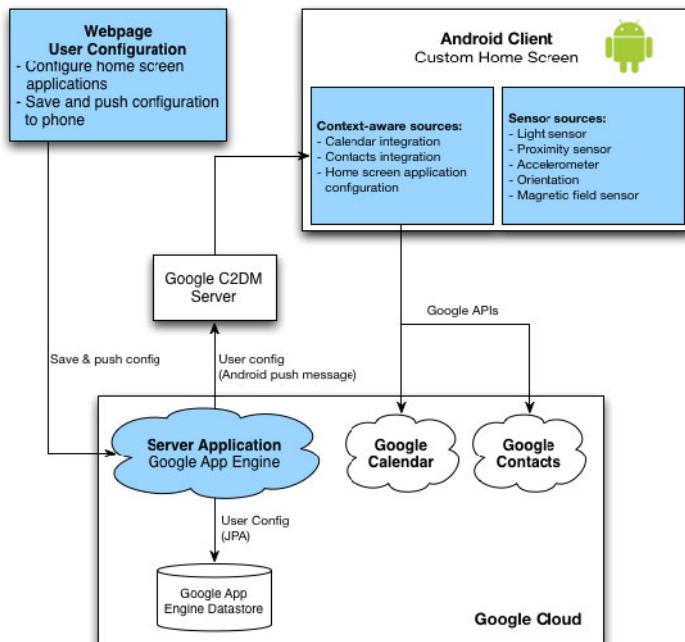


Fig. 2. System architecture

The blue boxes in the diagram represent the parts of the system we created. The white boxes, like Google calendar and Google contacts, are external systems we communicated with. The server application was deployed on Google App Engine and data stored in the Google cloud. We also integrated with the Google login service, making it possible for all users with a Google account to log on to the system.

When a new user logs in to the system with a Google account we automatically store this information in the Google App Engine data store. By using JPA (Java Persistence API) in combination with objectify we were able to store and retrieve data with minimal boilerplate code. The push functionality was implemented using the C2DM (Cloud to Device Messaging) framework. This is described in more detail in

the next section. Each C2DM enabled device would register with Google and receive a registration id. This id was sent to our cloud server through a normal HTTP GET call. We then proceeded to store this registration id in the data store. When the user pushed the save button on the webpage (see figure 1), to update the settings on the Android home screen, this would trigger a push message to be sent to the previously persisted registration id.

In addition to this we added support for input from several sensors available on the mobile device. In our experiment we look at input from 5 sensor sources: accelerometer, magnetic field, orientation, proximity and light. For light we added the functionality of changing the background color based on the value received from the sensor. If the room was dark the screen shows a white background. When the user moves to a room with more light the screen will gradually get darker.

4 Application Walkthrough

The user interface was evaluated with a sample of users who expressed their opinions in respect of statements targeting ease-of-use, functionality and ease of adaptation. Overall results indicated that the users have a strong positive bias towards the application and that the adaptation helped them to tailor the device to their needs by using the different context aware mechanisms. In this section we will provide an overview of the mobile Android client in light of user evaluation feedback. Figure 3, below, shows the configurable home screen as displayed on the device.

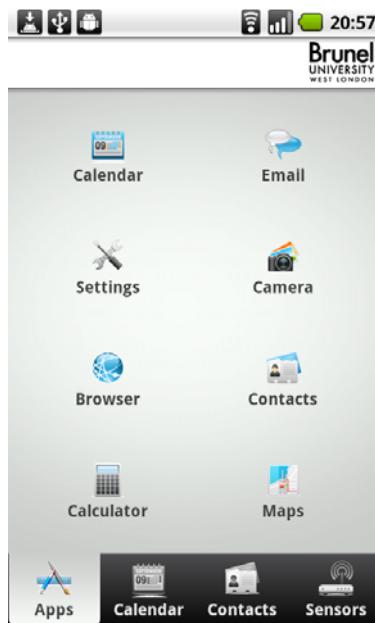


Fig. 3. Application home screen

The information presented on the home screen is quick launch icons for different applications installed. The specific information on the screen is tailored in accordance with the chosen user settings in the cloud based web interface (shown in figure 2). By taking advantage of the brand new Cloud 2 Device Messaging (C2DM) API provided by Google for Android 2.2 we were able to push messages to the devices without the application running. Not only is this a much simpler solution for the users it also has advantages like providing better battery life since we do not have to rely on a polling mechanism. The Android Cloud to Device messaging service is currently in beta and is not available to the general public. Feedback from user testing very positive to this feature and one commented on the possibility for this home screen view to replace the original Android phone home screen. We see this as a viable and interesting idea worthy future pursuit.

Also displayed in figure 3, in the bottom part of the picture, is one of the navigation menus from in the application. By selecting the “Calendar” tab, the application switches to calendar mode and presents upcoming appointments to the user. The upcoming appointments presented are only relevant to the users context, meaning that information are filtered based on a user-context algorithm. The special tags, \${type=work} and \${type=leisure}, provide one out of several foundations for the application to compute the users context. Other factors influencing the decision is context-aware dimensions such as GPS tracked location and time. Once a user-context is defined, this will also influence the selection of contacts displayed if selecting the next tab in the bottom part menu. The features of custom selected contacts and calendar items generated a lot positive feedback and response in our user test group. One user commented: “The tailoring and filtering of information eases my job of searching for information”. Another user pointed out that this feature could also be applied to other cloud stored data such as documents and notes.

In our application we have, as previously described, input from five different sensor sources: accelerometer, magnetic field, orientation, proximity and light. This makes for another novel dimension to use in our user context computation. We actively use sensor input for other tailoring of the application as well. Figure 4 and figure 5 displays an example of the light sensors in action. By integrating with the different sensors, such as light, we are able two automatically adjust brightness of the display for the end user. By continuously measuring the amount of light around the devices we are capable of changing the background color based on the value received from the sensor. This shows in the application as a darker background color in the application the brighter the room is. This adaptation is especially useful for the user when using the application in outdoor scenarios because of the improved readability of the user interface.



Fig. 4. Light sensor in action 1



Fig. 5. Light sensor in action 2

Figure 4 and figure 5 shows a sub screen of the implementation of the light sensor. The figures show how the background can change according to the measured lux value, also displayed in these screenshots for information purposes. From user testing the implementation of the light sensor as a context-aware dimension variable for user context computation where very welcomed. The ease of use of the application under very different lighting conditions due to the background color adaptation where highlighted as the major positive factor. As figure 6 illustrates we are currently looking into making even more sensor data available and include them into the context-aware equation.



Fig. 6. Sensor panel

5 Concluding Remarks

The most important aspect of our work is how a flexible and extensible solution can be integrated with the Google cloud to provide value to the users. We can take advantage of many different context sources and adapt the home screen to what the user wants. Our experiment gives a few examples, but there are many other possibilities that can be added to the system. Another important part of the system is the integration with sensors on the mobile device. Android gives the developer full access to read values from the sensors on the phone, this is an exciting feature that we would like to extend in future versions of the system.

Overall results indicated that the users have a strong positive bias towards the application and that the adaptation helped them to tailor the device to their needs by using the different context aware mechanisms.

References

1. Mei, L., Chan, W.K., Tse, T.H.: A Tale of Clouds: Paradigm Comparisons and Some Thoughts on Research Issues. In: Proceedings of the 2008 IEEE Asia-Pacific Services Computing Conference (APSCC 2008), pp. 464–469 (2008)
2. Weiser, M.: The computer for the 21st century. In: Human-computer interaction: toward the year 2000, pp. 933–940. Morgan Kaufmann Publishers Inc., San Francisco (1995)
3. Google Mobile: Android basics: Getting to know the Home screen (2010),
<http://www.google.com/support/mobile/bin/answer.py?answer=168445#1149468> (last visited October 5, 2010)
4. Kawaguchi, K.: Hudson Extensible continuous integration server (2010),
<http://wiki.hudson-ci.org/display/HUDSON/Meet+Hudson> (last visited October 10, 2010)
5. Göker, A., Watt, S., Myrhaug, H.I., Whitehead, N., Yakici, M., Bierig, R., Nuti, S.K., Cumming, H.: An ambient, personalised, and context-sensitive information system for mobile users. In: Proceedings of the 2nd European Union symposium on Ambient intelligence, pp. 19–24. ACM, Eindhoven (2004)
6. Dey, A.K.: Understanding and using context. *Journal of Personal and Ubiquitous Computing* 5(1), 4–7 (2001)
7. Parviainen, M., Pirinen, T., Pertilä, P.: A Speaker Localization System for Lecture Room Environment. *Machine Learning for Multimodal Interaction*, 225–235 (2006)
8. Chatzigiannakis, I., Kinalis, A., Nikoletseas, S.: Priority Based Adaptive Coordination of Wireless Sensors and Actors. In: Proceedings of the 2nd ACM International Workshop on Quality of Service & Security for Wireless and Mobile Networks (2008)
9. Boulis, A., Han, C., Srivastava, M.: Design and Implementation of a Framework for Efficient and Programmable Sensor Networks. In: Proceedings of the 1st International Conference on Mobile Systems (2003)
10. Mei, L., Chan, W.K., Tse, T.H.: A Tale of Clouds: Paradigm Comparisons and Some Thoughts on Research Issues. In: Proceedings of the 2008 IEEE Asia-Pacific Services Computing Conference, pp. 464–469. IEEE Computer Society, Los Alamitos (2008)
11. Mowbray, M., Pearson, S.: A client-based privacy manager for cloud computing. In: Proceedings of the Fourth International ICST Conference on Communication System Software and Middleware, pp. 1–8. ACM, Dublin (2009)
12. Bilandzic, M., Foth, M., Luca, A.: CityFlocks: Designing Social Navigation for Urban Mobile Information Systems. In: Proceedings ACM Designing Interactive Systems (2008)
13. Rodden, T., Cheverest, K., Davies, K., Dix, A.: Exploiting context in HCI design for mobile systems. In: Workshop on Human Computer Interaction with Mobile Devices (1998),
<http://www.dcs.gla.ac.uk/~johnson/papers/mobile/HCIMD1.html>
14. Ludford, P., Rankowski, D., Reily, K., Wilms, K., Terveen, L.: Because I carry my cell phone anyway: functional location-based reminder applications. In: Proceedings of Conference on Human Factors in Computing Systems, April 2006, pp. 889–898 (2006)
15. Dey, A.K., Abowd, G.: Towards a Better Understanding of Context and Context-Awareness. In: Proceedings of the 1st International Symposium on Handheld and Ubiquitous Computing (1999)

Evaluation of User Support of a Hemispherical Sub-display with GUI Pointing Functions

Shinichi Ike^{1,*}, Saya Yokoyama¹, Yuya Yamanishi¹, Naohisa Matsuuchi¹, Kazunori Shimamura², Takumi Yamaguchi¹, and Haruya Shiba¹

¹ Department of Electrical Engineering and Information Science, Kochi National College of Technology, 200-1 Monobe-Otsu, Nankoku Kochi, Japan

² School of Information, Kochi University of Technology, 185 Tosayamadacho-Miyanouchi, Kochi, Japan

s1001@gm.kochi-ct.jp,
shiba@ee.kochi-ct.ac.jp

Abstract. In this paper, we discuss the effectiveness of a new human interface device for PC user support. Recently, as the Internet utilization rate has increased every year, the usage of PCs by elderly people has also increased in Japan. However, the digital divide between elderly people and PC beginners has widened. To eliminate this digital divide, we consider improving the users' operability and visibility as our goal. We propose a new hemispherical human-computer-interface device for PCs, which integrates a hemispherical sub-display and a pointing device. Then we evaluate the interface device in terms of its effectiveness of operability and visibility. As seen from the analyses of a subjective evaluation, our interface device obtained good impressions results for both elderly people and PC beginners.

Keywords: PC User Support, Human-Computer-Interface Device, GUI Pointing Device, Hemispherical Display.

1 Introduction

As of 2010, the proportion of elderly people who are at least 65 years old is over 23% in Japan, and it is predicted that in the future, the percentage will increase with decrease in the population of Japan [1]. The number of elderly personal computer (PC) users will increase with the increase in the Internet utilization rate, which has been increasing every year [2]. However, elderly people lack easy access to information because of the loss of motor functions and sight due to aging. For example, PC displays can provide more information per unit area because they now have higher resolution. However, the problem is that the characters and icons on the display are very small and indecipherable by elderly people. Moreover, it is difficult for elderly people with loss of motor functions to move the mouse pointer and to click the mouse switches. Hence, a difference in the ability to access information

* Corresponding author.

(digital divide) arises between the users who can control a mouse pointer and those who cannot. In particular, the digital divide between elderly people and PC beginners has widened.

As a method to eliminate this digital divide, we consider improving the users' operability and visibility for a GUI system as our goal. In this study, we examined a new hemispherical human-computer-interface device (hemispherical HCID) for PCs that supports operability and visibility for elderly people and PC beginners.

2 Outline of Hemispherical HCID

We propose a new hemispherical HCID that integrates a hemispherical sub-display and a pointing device. Our HCID is placed near the user, at a location closer than the main PC display. The user can move the mouse pointer by directly touching the HCID. Placing his/her hand on the dome body and tilting it in the direction of movement, enables the movement of the mouse pointer. The hemispherical sub-display shows the PC desktop image, which is extracted around the mouse pointer and enlarged. The enlarged image follows the movement of the mouse pointer.

The work area of the main PC display is not limited by the hemispherical sub-display. Our HCID users usually operate the device while looking at the main PC display. If they have difficulty in recognizing the characters or icons on the main display, they use the hemispherical sub-display at hand. The users do not lose the information of the entire work area, because they can view the sub-display at their convenience. This manipulation method is a characteristic of the hemispherical HCID that integrates display and pointing function. Moreover, users do not lose the location of the mouse pointer, because the pointer is always visible on the sub-display.

We believed that the spherical form would be easy for users with physically disabled fingertips to operate, and that elderly people can operate the device similar to patting their grandchildren on the head. Earlier, users had to move the device on the desk while holding it. Our HCID puts less strain on the users' shoulders and arms, so that the physical strain of users can be reduced. A device with a sub-display and pointing function achieves support for both operability and visibility.

3 Previous Study

In our previous study, a pointing device with a flat liquid crystal display (LCD) in a transparent hemisphere-dome bottom, called "OPR-LENS device" [3] was built as a prototype and its user evaluation was performed. In this evaluation, we determined its operability as a pointing device and its effectiveness as a visibility support [4, 5]. Especially, for elderly people using the OPR-LENS system, it was found that it took a little time to operate; however, as compared to the use of a mouse, the error rate decreased and steady pointing was confirmed [5]. Moreover, from the evaluation of a fish-eye-processed image, it is also clear that this image has better quality compared with the one in which the original image was simply expanded.

To improve the effectiveness of visibility support, we changed the flat display into a hemispherical display. This is expected to improve user visibility, because the hemispherical display can widen the display surface area in comparison with the flat one, and can obtain an enlarged image similar to a fish-eye-processed image.

Previously, to build the hemispherical display, we proposed a two-flow system that integrated two or more optical lenses. These lenses were assembled on a flat LCD, and the LCD's image was projected onto a hemispherical surface by the refraction and reflection of light. The lenses of the two-flow system were designed by a ray tracing method, and we performed a numerical simulation of the visual feature [6]. As a result of prototyping, however, it was concluded that with the two-flow system it was difficult to project the PC desktop image [7]. Theoretically, the attenuations of light intensity are 1/2 and 1/20 at the LCD and optical lenses, respectively [6]. We verified these values using a luxmeter. It was believed that a high-intensity LED backlight could improve the lower luminance at the optical lenses and the LCD. However, the attenuation was indeed beyond our expectations. The attenuations of light intensity were 1/10 and 1/12 at the LCD and optical lenses, respectively. The brightness of the LED backlight could be increased only to a certain extent because of the generation of heat. Then, we create a new prototype of the structure built in a micro-projector. A structure built in a micro-projector has some advantages. For example, the structure is simpler than a two-flow system, and a high-intensity light source can be easily obtained.

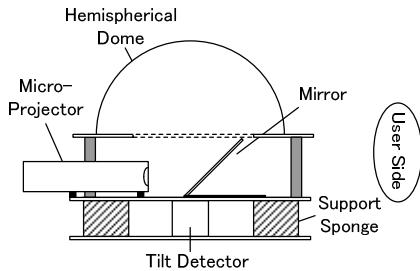
Table 1 shows the comparison of the average luminance on the surface of the dome between the two-flow system and the system that was built in a micro-projector. It was assumed that the image was projected to occupy 60% of the surface of the dome. In both systems, it was assumed that the luminance of the light source is 10 lumen. Theoretically, we found that the system with a micro-projector improves the surface luminance by about 4% compared with the previous system. In general, the luminance of the PC display is lower than about 500 lux, and therefore, it is not a problem to change to a system built in a micro-projector.

Table 1. Comparison of surface luminance

Projection System	Value [lux]
LCD + Optical Systems	453
Micro-projector	472

4 Structure of Hemispherical HCID System

A hemispherical HCID is mainly composed of a sub-display component and a user-operation-detection. Figure 1 shows the inner structure of the device. Figure 2 shows the image of the prototype of the device. The output of the magnifier tool of the Windows OS standard functions is used as the sub-display image that is projected onto a dome surface. The game controller for the USB port is used as the operation detector. We also used the JoyToKey (version 4.5.3) software, which converts the controller input into mouse input.

**Fig. 1.** Inner structure of hemispherical HCID**Fig. 2.** Image of the prototype of hemispherical HCID

5 Subjective Evaluation of Hemispherical HCID

In the previous study, a user evaluation of the OPR-LENS device that assembles a flat LCD in the transparent-dome bottom was performed. However, the device with the hemispherical image has not yet been examined, and it is necessary to examine its effectiveness in terms of operability and visibility. We built the prototype and performed a subjective evaluation. In this chapter, we explain the evaluation method and its results.

5.1 Experimental Methods

The evaluation subjects were 18 men and 44 women, a total of 62 individuals, who were selected regardless of their PC experience. Their ages ranged from 13 to 82: 27 from 10–19, 5 from 20–29, 7 from 30–39, 7 from 40–49, 16 over 50. First, we explained both the purpose of the device and its methods of operation to the subjects. After the subjects experienced the device, we performed a subjective evaluation of its operability and visibility by a categorical rating.

The questionnaires evaluated subjective impressions using a five-point rating scale: Better = 5, Slightly better = 4, Fair = 3, Slightly worse = 2, and Worse = 1. The questionnaires included the following five questions:

Question 1: Did you understand what the device is like?

Question 2: Is the sub-display at hand easily visible?

[Multiple answers allowed]

- (1) *It is easily visible because of the sub-display at hand.*
- (2) *An enlarged hemispherical image is easily visible.*
- (3) *The image is bright and easily visible.*
- (4) *The image is dark and difficult to see.*
- (5) *Only the image around the mouse pointer is good.*

Question 3: Is the device easy to manipulate?

[Multiple answers allowed]

- (1) *The shape of the hemispherical dome is easy to manipulate.*
- (2) *The shape of the hemispherical dome is difficult to manipulate.*
- (3) *It is easy to manipulate it because of the enlarged image at hand.*
- (4) *The enlarged image is unnecessary to have at hand.*
- (5) *This device gave you less strain on shoulders and arms.*
- (6) *This device gave you more strain on shoulders and arms.*

Question 4: Do you think this device is useful?

Question 5: If input devices such as this device are marketed, would you like to use them?

The hemispherical HCID was assembled near the right side of the laptop PC, and the subjects sat on a chair. For this experiment, we used a laptop PC whose display is a 10.4 in LCD with a resolution of 1024 × 768 pixels (XGA).

5.2 Results

We divided the subjects into two groups. The group, called Group A, spends <2 h per day on the computer; the other group, called Group B, spends >2 h per day on the computer. Table 2 shows the mean scores and standard deviation (SD) calculated for each group by age.

We conducted a t-test of statistical significance to compare the questionnaire results between Group A and Group B. The mean scores obtained in Question 1 were over 4 points for all ages, and there was no significant difference in terms of computer time at a significant level of $p < 0.05$. In Question 2, the subjects of Group A answered with a higher score compared with the subjects of Group B in the 10–19 and over-50 age groups, but there was no significant difference in terms of computer time. In Question 3, the subjects of Group A answered with a higher score compared with the subjects of Group B in the over-50 age group, and there was a significant difference in terms of computer time at a significant level of $p < 0.05$. In Question 4, the subjects of Group A answered with a higher score compared with the subjects of Group B in the 10–19 and over-50 age groups. Especially in the over-50 age group, there was a significant difference in terms of computer time at a significant level of $p < 0.05$. In Question 5, the subjects of Group A answered with a higher score compared with the subjects of Group B in the 10–19 and the over-50 age groups. There was no significant difference in their computer time.

Figures 3 and 4 show the comparison of the mean scores for each group based on age groups. We also conducted a t-test of statistical significance to compare the questionnaire results between the age groups in the same group. In Question 2, for Group A, the mean score for the 10–19 age group was the highest of all, and there were significant differences between the 10–19 and the other age groups at a significant level of $p < 0.05$. In Question 2, for Group B, the mean score in the over-50 age group was the lowest of all. In Question 3, for Group A, the mean score in the 10–19 and over-50 age groups was higher than in the 20–49 age group, and there were significant differences between them at a significant level of $p < 0.05$. In Question 3, for Group B, the mean score in the over-20 age group was lower than three points. In Questions 4 and 5, for Group A, the mean score in the 10–19 age

group was the highest of all, and the mean score in the 20–49 age group was the lowest of all. In Questions 4 and 5, for Group B, the mean score in the over-50 age group was lower than that of the under-40 age group.

Table 2. Comparison of the results of the questionnaire between two groups (<2 h, >2 h)

Age	Question number	Group A (<2 h)		Group B (>2 h)		p-value
		Mean score	S.D.	Mean score	S.D.	
		Sample size = 20		Sample size = 7		
10–19	1	4.50	0.69	4.71	0.49	0.457
	2	4.20	0.70	4.00	1.00	0.564
	3	3.65	0.93	3.57	1.40	0.868
	4	4.45	0.76	3.86	0.90	0.102
	5	3.70	0.86	3.14	1.57	0.401
		Sample size = 9		Sample size = 10		
20–49	1	4.22	0.67	4.70	0.48	0.097
	2	2.67	1.22	3.50	1.08	0.133
	3	2.56	0.88	2.40	0.84	0.700
	4	3.44	1.01	3.80	1.03	0.460
	5	2.44	0.73	3.10	1.60	0.263
		Sample size = 11		Sample size = 5		
>50	1	4.64	0.50	4.60	0.55	0.898
	2	3.09	1.14	2.20	1.30	0.186
	3	3.73	1.10	2.20	1.30	0.029
	4	4.00	1.00	2.40	1.34	0.018
	5	2.91	1.58	1.80	1.30	0.193

Table 3 shows the results of the impressions of the evaluation in Questions 2 and 3. We observed that the 10–19 and over-50 age groups' subjects had better impressions compared with the 20–49 age group in (1), (2), (3), and (5) in Question 2. However, the brightness of the sub-display appeared dark to the over-50 age group. In Question 3, we obtained the results that the 10–19 and over-50 age groups' subjects had better impressions compared with the 20–49 age group in (1), (3), and (5). However, it seemed that the subjects in the 20–49 age group felt that the enlarged image was not needed at hand.

From the results of Tables 2 and 3 and Figures 3 and 4, it is considered that our device is effective for operability and visibility, especially for the 10–19 and the over-50 age groups of Group A (<2 h of computer time). Thus, the user support provided by our device was effective for elderly people and PC beginners.

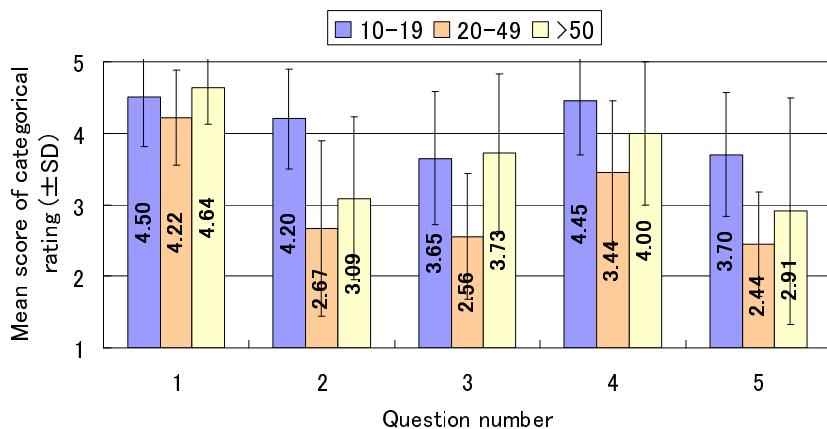


Fig. 3. The comparison of the results of Group A (<2 h) by age

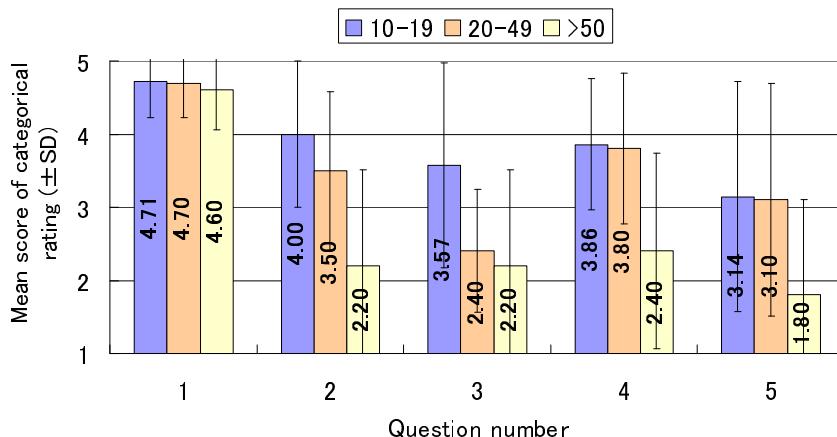


Fig. 4. The comparison of the results of Group B (>2 h) by age

Table 3. The results of the impressions of visibility and operability

Items	Frequency / Sample size [%]		
	10–19 Sample size = 27	20–49 Sample size = 19	>50 Sample size = 16
Question 2			
(1) It is easily visible because of the sub-display at hand.	67	7	31
(2) An enlarged hemispherical image is easily visible.	30	16	31
(3) The image is bright and easily visible.	30	0	6
(4) The image is dark and difficult to see.	30	26	56
(5) Only the image around the mouse pointer is good.	41	5	25
Question 3			
(1) The shape of the hemispherical dome is easy to manipulate.	48	16	31
(2) The shape of the hemispherical dome is difficult to manipulate.	22	21	25
(3) It is easy to manipulate it because of the enlarged image at hand.	56	5	50
(4) The enlarged image is unnecessary to have at hand.	4	11	6
(5) This device gave you less strain on shoulders and arms.	59	11	31
(6) This device gave you more strain on shoulders and arms.	0	0	6

6 Related Works

In this chapter, we introduce a variety of spherical displays or spherical-shaped human-computer-interface devices and compare our device with the related works.

- **Geo-Cosmos [8]**

Geo-Cosmos is a spherical display of about 6.5 meters in diameter and about 15 tons in weight that has been displayed in the National Museum of Emerging Science and Innovation. 3715 panels that arrange 256 LEDs cover the entire sphere.

- **Tangible Earth [9]**

Tangible Earth is a spherical display of about 1.2 meters in diameter and about 60 kilograms in weight. An entire image is projected from the inside of the sphere by

a projector with a fish-eye lens. The sensor detects the movement when its body is turned by hand, and a new earth image that is calculated instantly by the computer is projected. The touch panels cover the surface.

- **Magic Planet [10]**

Magic Planet is a spherical display of about 0.4–3 meters in diameter. The image is projected from four projectors onto the sphere.

- **I-ball 2 [11]**

The device integrates a transparent sphere display and a trackball interface. An LCD, a mirror, and a Fresnel lens that are assembled at the bottom of the ball form the image inside a transparent ball of 20 centimeters in diameter. The image appears to be suspended in the air, and the ball is integrated with an image that the users cannot touch, even if they want to. The transparent ball can be freely rotated like a trackball. This rotation is detected by an optical sensor, which enables an intuitive operation of the image.

- **Sphere [12]**

Sphere is an interactive spherical-display prototype. The device is built on a commercially available globe-projection system. Touch-sensing capabilities with an infra-red camera that shares the optical path with the projector are added.

Because we assume that our device is sized to be assembled on a desk, the system consisting of a covering LED on the surface, such as Geo-Cosmos and the system of projecting from outside, such as Magic Planet, are too large, and these systems are not suitable for our device. Tangible Earth, I-ball 2, and Sphere are used by rotating the sphere body. Our device uses an operation method different from that used by trackball.

7 Conclusions

In this study, we developed a hemispherical HCID to support the operability and visibility for elderly people and PC beginners who are prone to information shortfalls. We built a prototype and performed an evaluation to examine the effectiveness of user support by the hemispherical HCID. We conducted a t-test of statistical significance to compare the questionnaire results between the subjects with <2 h of computer usage per day (Group A) and those with >2 h of computer usage per day (Group B). In addition, we also conducted a t-test to compare the questionnaire results between the age groups in the same group. According to the analyses of the evaluation results, it is considered that our device is effective in terms of both operability and visibility, especially for the 10–19 and over-50 age groups in Group A (<2 h of computer time). Thus, we concluded that the user support provided by our device was effective for elderly people and PC beginners.

However, some problems were pointed out by the subjects. For instance, the size of the dome was too large, the brightness and the contrast of the surface image should be higher, and the operation sensitivity should be improved. If we can overcome these problems, it is possible to enhance the effectiveness of user support for all users. We would like to improve our device and propose other new devices for PC user support.

Acknowledgments. We thank Electric Parts Kochi Co., Ltd., Perori bakery, and Tobigaike junior high school for their cooperation in the category appraisal. This study was partially supported by a Grant-in-Aid for Scientific Research (B, Project No. 22300300 and C, Project No. 20500119).

References

1. Japan Cabinet Office: White paper on aging society for 2010,
<http://www8.cao.go.jp/kourei/whitepaper/w-2010/zenbun/22index.html>
2. Ministry of Internal Affairs & Communications: Information & Communications Statistics Database, <http://www.soumu.go.jp/johotsusintohei/whitepaper/ja/h22/index.html>
3. Shiba, H., Yamaguchi, T., Shimamura, K.: A proposal of a hemispherical display device having a built-in pointing device for a partially enlarged GUI display. In: Proc. of IWAIT 2006, pp. 508–513 (2006)
4. Yamaguchi, T., Shiba, H., Suga, M., Shimamura, K.: Evaluation of an Operational Performance in the Pointing Device ‘OPR-LENS’ manipulated by a Stroking Operation. In: Proc. of HIS 2004, pp. 1175–1178 (2004)
5. Koji, M., Yamaguchi, T., Shimamura, K., Shiba, H.: Method for Evaluating Effects of Visual Support Provided by New Pointing Device with Magnified Display. In: APCHI 2006, vol. 37 (2006)
6. Kawada, M., Shiba, H., Yamaguchi, T., Shimamura, K.: Examination and production of a hemispherical display. In: Proc. of FIT 2006, K-066, pp. 527–528 (2006)
7. Ike, S., Shiba, H., Yamaguchi, T., Shimamura, K.: Influence of Hemispherical Display on Information Equipments Operation by Elderly User and Beginner User. In: Proceedings of International Workshop on Information Technology (iwit 2009), pp. 101–104 (2009)
8. Geo -Cosmos, <http://www.miraikan.jst.go.jp/exhibition/geo-cosmos/>
9. Tangible Earth, http://www.tangible-earth.com/index_ja.html
10. Global Imagination, <http://www.globalimagination.com/>
11. Ushida, K., Harashima, H., Ishikawa, J.: i-ball 2: A Crystal-Ball-Like Display Observable by Multi Users: Interactive Applications Using Track Ball Interface. The report of technical research of The Institute of Electronics, Information and Communication Engineers (IEICE), MVE, pp. 15–20 (2003)
12. Sphere: A Multi-Touch Interactive Spherical Display,
<http://research.microsoft.com/en-us/um/people/benko/projects/sphere/>

Uni-model Human System Interface Using sEMG

Srinivasan Jayaraman¹ and Venkatesh Balasubramanian²

¹ TCS Innovation Labs –Bangalore, TATA Consultancy Services, Bangalore - 560 066, India

² Rehabilitation Bioengineering Group, Department of Engineering Design, IIT-Madras,
Chennai -600036, India

Srinivasa.j@tcs.com, chanakya@iitm.ac.in

Abstract. Today's high-end computer systems contain technologies that only few individuals could have imagined a few years ago. However the conscious input device ergonomics design is still lagging; for example, the extensive usage of computer mouse results in various upper extremity musculoskeletal disorders. This endower towards the developed of HSI system, that act as an alternative or replacement device for computer mouse; thereby one could avoid musculoskeletal disorders. On the other hand, the developed system can also act as an aid tool for individuals with upper extremity disabled. The above issue can be addressed by developing a framework for Human System Interface (HSI) using biological signal as an input signal. The objective of this paper is to develop the framework for HSI system using Surface Electromyogram for individuals with various degrees of upper extremity disabilities. This framework involves the data acquisition of muscle activity, translator algorithm that analysis and translate the EMG as control signal and a platform independent tool to provide mouse cursor control. Thus developed HSI system is validate on applications like web-browsing, simple arithmetic calculation with the help of GUI tool designed.

Keywords: EMG, HSI, Computer Cursor Control.

1 Introduction

Growth in silicon provide pathway toward increase in Communication between human seems usually much simpler than one involve in human machine interaction. This became more intricate when a person is disabled. On the other hand, the conscious input device ergonomics is still lagging; example, extensive usage of computer mouse produces various upper extremity musculoskeletal disorders. Over few decades, various research groups (Ten Kate *et.al.*, 1983/1984, LaCourse R.J. and Hludik, Jr. C. F. 1990, Naito, *et.al.*, 2002, Hori.j., *et.al.*, 2006) have begun work, to solve the above issues and aid communication and mobility of disable individuals and elderly individuals as well. These research works on alternative or aid techniques focus on increase the quality of life and allowing them a more autonomous and independent lifestyle and greater chances of social integration. One such communication aid technology is the Human system interface (HSI) or human computer interface (HCI) and brain computer interface (BCI).

Research on alternative or aid techniques of computer control has focused on three types of body functions: speech, gustier and bioelectrical signal. These aid techniques vary by themselves by the command signal provided to control the peripheral devices, translation algorithm or display device. This alternative communication is achieved by using biofeedback system. Common modalities of Human system interface system (HSI's) are based on hand movements; because human hand allows accurate selection, positioning and appropriate force and acceleration can be applied with the help of visual feedback mechanism. Thus, hand movement is exploited in the design of numerous interface devices—keyboard, mouse, stylus, pen, wand, joystick, trackball, etc.

One such approach is, HSI using muscle activities for computer cursor control system will get input from human physiological signal, this physiological input is an electrical signal acquired from upper extremity muscle movement. There are several approaches to measure these signals; surface Electromyogram (sEMG) (Basmajian, J.V., and DeLuca C.J., 1985) to collect the corresponding human muscle movement. Further, the origin and characteristics of EMG and interpretation can be found in Rau, Schulter and Disselhorst-Klug, 2004; Soerberg and Knutson, 2000; Stegeman *et.al.*, 2000; Burgar and Rugh, 2000 and Boxtel and Schomaker, 1983.

Over the years, Electromyogram (EMG) has been used to diagnose and treat neuromuscular disorders and to demonstrate improvements in muscle functioning, following conventional treatment or neuromuscular stimulation. Later on EMG was used as an active component of biofeedback system, not only for monitoring but also as a control parameter (Malmo and Malmo 2000, Barniv *et al.*, 2005, Barreto *et al.*, 1999 and 2000, Gregor *et al.*, 1991, Harver, Segreto and Kotses, 1992) and EMG has become an important tool in rehabilitation.

Kennedy et al., (2000) presents evidence that cursor control is achieved by implanted electrode, where EMG signals were used as control signals. In this system increase or decrease of firing rate is used for X-Y control of cursor. Later, continues training the users for activating the cursor and name it as cursor cortex. In another studies by Harver *et al.*, (1992) reported that EMG is used in biofeedback control for stability of muscle as a rehabilitation tool. Trejo, 2003 report that, EMG signal are applied as a control signal for an aircraft simulator and keypad typing in virtual keyboard. Trejo's Hidden Markov model for training the EMG signals for various application. Kwon and Kim (1999), developed a low cost wireless mouse for disabled, still this system have limitation due to practice of virtual lead positioning, which leads to accuracy problem.

With these issues in mind, this paper aims to build an inexpensive, self adaptive HSI system using EMG signal as control comment. The main motivation of this work is that computers should be adapting to people and not the other way, even for disabled.

2 Protocol

The uni-modal HSI system utilizes the sEMG signal as input source to control the computer cursor control is known as sEMG based HSI system. Prior to the start of

the experiment the subjects were given verbal cues to perform specific types of hand movements, to move the cursor in horizontal and vertical direction. Initial delay of 2 sec is given for the system; this delay was given to ensure the present status of the system. As subject starts to contract the muscle in order to achieve or reach their target, sEMG amplifier acquires the signals from the resultant connected electrode. The acquired signal is digitized and further processed is performed to extract the feature parameters as explained early. Each and every muscle group has its own reference threshold. The present experimental setup is divided into three parts:

1. Acquiring EMG signal and feature parameter extraction,
2. Mapping algorithm based on extracted features and transmit the control signal
3. Receiving the control signal and performing the corresponding cursor movements using mapping algorithm.

Raw surface Electromyogram (sEMG) signals were acquired using bio-amplifier from the corresponding voluntary agonistic muscle movement. Acquired raw sEMG signals were pre-processed and feature parameters were extracted. Translator algorithm or classification of signal can be performed by two techniques

- a) Frequency estimation method and
- b) Model based classification.

In the present work, we have adapted frequency estimation techniques. The control signals are communicated through parallel port to the second computer. The second system will receive the signal and a mapping algorithm is performed. Based on this translator output the next step of cursor will be determined. The acquisition of sEMG, processing and receiving technique were implementation of real time cursor control using sEMG.

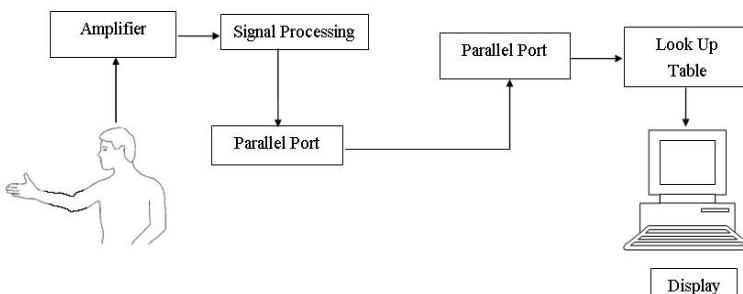


Fig. 1. Schematic block diagram representation of unimodal HIS system used for computer cursor control using sEMG as control signal

Developed uni-model system was evaluated over three different applications: a) virtually simulated calculator environment for doing arithmetic operation, b) virtually simulated web surfing, and c) motor control, where the direction and speed of dc motor and stepper motor was controlled.

3 Methodology

3.1 Subject

Ten subjects with mean age of 26.2 years (range 23 – 29 years) and mean weight of 60.17 Kg (range 45 – 68 Kg) participated in the study, which involved cursor movement for 10 minutes.

3.2 Electrode Placement

Muscle group considered for this studies were Brachio radialis (BR), Biceps brachii (BB), Flexor carpus radialis (FCR), Triceps brachii (TB). Figure 2 shows the placement of Ag-AgCl electrode on the subjects for the developed HSI system.

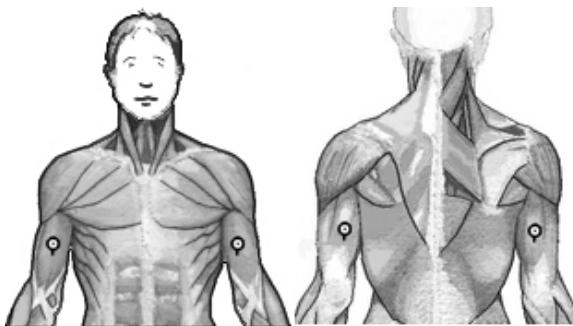


Fig. 2. Electrode placement diagram for recording sEMG signals

3.3 System Implementation

Functional block diagram of uni-modal HSI System based on frequency estimation method is shown in Figure-3. Raw sEMG signals were filtered using second-order Butterworth filter with a pass band of 20 – 400 Hz. AC line interferences were eliminated with a second order notch filter with a band stop of 47 – 51 Hz. Filtered signals were rectified with respect to mean, and sEMG signals were analyzed in frequency domain using power spectrum. The power spectrum shows the total power of the signal analyzed along with the frequency spectrum. This PSD estimation is obtained by calculating the discrete Fourier transform (DFT) on sEMG data digitized at 1000Hz, after applying Hamming window. Hamming–windowed algorithm was applied on rectified sEMG signals and feature parameters were extracted.

The mean power frequency (MPF) is used as a feature parameter to control the cursor position. The MPF of sEMG is calculated using Equation 1. Individual muscle group has its own reference threshold value and based on the threshold level comparison, the new step of cursor position is decided as shown in Table 1.

$$\text{MPF} = \left[\frac{f_1 \times p_1 + f_2 \times p_2 + \dots + f_n \times p_n}{p_1 + p_2 + \dots + p_n} \right] \quad (1)$$

Where n =1, 2, 3....etc

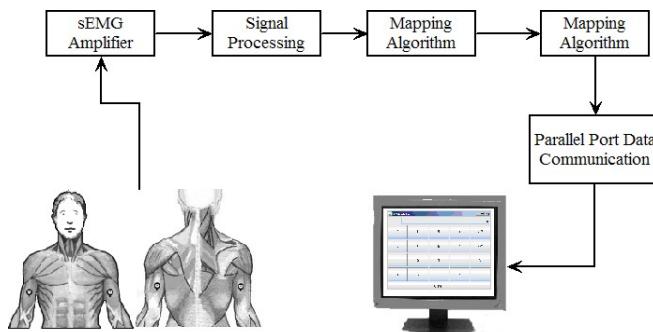


Fig. 3. Functional block diagram of uni-modal HSI System based on frequency estimation method

Table 1. Relation between Cursor Action, Threshold and Muscle Group

Muscle group	Threshold	Classification Algorithm Output	Computer Cursor Action
		0	No Action
Brachio radialis	TH 1	1	Right cursor movement
Biceps brachii	TH 2	2	Left cursor movement
Flexor carpus radialis	TH 3	3	Up cursor movement
Triceps brachii	TH 4	4	Down cursor movement

The new cursor position is determined based on the thresholds value, which is a sign of the corresponding contraction muscle group activity. Thus generated control signal is transferred to the translator unit for performing the mapping algorithm. The communication between the translator and mapping algorithm is performed via parallel port of computer. This control signal transmissions were represented by 4-bit equivalent. All these signal processing and control signal transmissions is performed with Matlab™.

Transmitted control signals were acquired from parallel port and decision logic is applied. The control signals are mapped with the mapping algorithm and with respect to the decision logic corresponding cursor actions will be executed. Based on the mapping algorithm the mouse cursor movement is performed. Up movement is performed when the received input signals were 1, it means that TH1 > TH2. Down movement is performed when the received input signals were 2, which means that TH1 < TH2. The mapping algorithm and corresponding cursor action execution is performed with Java™ as shown in Table 2.

4 Result and Discussion

The frequency based sEMG HSI's performance was evaluated on healthy volunteers as online experiment and it is 90.2% with a maximum correct classification percentage of 92%. Reduction in system performance was due to the synchronization between users and system. Since it is a 4-bit representation, the possible number of control data transmitted is 2^n (i.e. 16), however for this experimental protocol we have used only 4 different cursor movement actions, such as right, left, up and down movement. The other different actions are left for future system extension depending on the user stipulation. A simulated calculator using java and web surfing applications are realized and tested on healthy individuals. The feature parameters were selected with cautious knowledge to achieve the performance and accuracy of the HSI's. The performance of the system was measured using Equation 2 and there efficiency is as shown in Table 3.

$$\text{Accuracy} = \frac{\text{Correct output}}{\text{Total number of input}} \times 100[\%] \quad (2)$$

Table 2. Correct/Incorrect Classification Percentages

No. of Subject	Classification Percentage (%)	
	Correct	Incorrect
1	89.1	10.9
2	92.05	7.95
3	89.43	10.57
4	90.2	9.8
5	89.13	10.87
6	89	11
7	91	9
8	89	11
9	90	10
10	92	8
Average	90.19	9.81

Thus developed HSI system using sEMG as input signal environment or interface design affects the subject comfortless with the existing HSI system and possibly the information transfer rate. To improve the accuracy and efficiency of the system, synchronization between user and computer is important; this lead to embracement of 2 sec delay in the system. The influx delay affects the performance of the system and information transfer rate. Off-line acquisition and analysis of biosignal were used to arrive at an optimal setup of control parameter and feature extraction.

5 Conclusion

The influence of computers in society increases the limitations of current human system interface. Current HSI system restricts the information and command flow between the user and the computer system. They are, for the most part, designed to be used by a small group of experts who have adapted themselves to the available HSI systems.

In this work, a novel algorithm and model is proposed to overcome the problem of one to one mapping between frequency vs action, and native feel vs robustness of system. Thus developed HSI's would serve as assisting device for various types and degrees of disabled individuals.

Acknowledgements. Authors wish to thank all members of Rehabilitation Bioengineering Group (RBG) at IIT Madras and the volunteers who participated in this study Text for this section. The authors would like to thank IIT Madras for the support in the form of basic infrastructure and graduate assistantship provided to the first author.

References

1. Barniv, Y., Aguilar, M., Hasanbelliu, E.: Using EMG to anticipate head motion for virtual-environment applications. *IEEE Transactions on Biomedical Engineering* 52(6), 1078–1093 (2005)
2. Barreto, A.B., Scargle, S.D., Adjouadi, M.: A Real-Time Assistive Computer Interface for Users with Motor Disabilities. *SIGCAPH Newsletter* (64), 6–16 (1999)
3. Barreto, A.B., Scargle, S.D., Adjouadi, M.: A Practical EMG-based Human-Computer Interface for Users with Motor Disabilities. *Journal of Rehabilitation Research And Development* 37(1), 53–63 (2000)
4. Boxtel, V.A., Schomaker, B.R.L.: Motor unit firing rate during static contraction indicated by the surface EMG power spectrum. *IEEE Transactions on Biomedical Engineering* BME30(9), 601–609 (1983)
5. Burgar, G.C., Rugh, D.H.: An EMG Integrator for muscle activity studies in ambulatory subjects. *IEEE Transactions on Biomedical Engineering* BME30(1), 66–69 (2000)
6. Basmajian, J.V., DeLuca, C.J.: *Muscles alive: their functions revealed by electromyography*. Williams and Wilkins, Baltimore (1985)
7. Gregor, R.J., Broker, P.J., Ryan, M.M.: The biomechanics of cycling. *Exercise and Sport Sciences Reviews* 19, 127–169 (1991)
8. Harver, A., Segreto, J., Kotses, H.: EMG stability as a biofeedback control. *Biofeedback and self-regulation* 17(2), 159–164 (1992)
9. Hori, J., Sakano, K., Saitoh, Y.: Development of a communication support device controlled by eye movements and voluntary eye blink. *IEICE Trans. Inf. & Syst.* E89-D(6) (2006)
10. Malmo, R.B., Malmo, H.P.: On electromyographic (EMG) gradients and movement-related brain activity: significance for motor control, cognitive functions, and certain psychopathologies. *International Journal of Psychophysiology* 38, 143–207 (2000)
11. Naito, A., Nozawa, A., Tanaka, H., Ide, H.: Communication substitution system using the eye blink and eye movement. In: Proc. FIT, pp. 457–458 (2002)

12. Kennedy, P.R., Bakay, R.A.E., Moore, M.M., Adams, K., Goldwaith, J.: Direct control of a computer from the human central nervous system. *IEEE Transaction of Rehabilitation Engineering* 8(2), 198–202 (2000)
13. Kwon, H.S., Kim, C.H.: EOG –based glasses type wireless mouse for the disabled. In: *Proceedings of the First Joint BMES/EMBS Conference Serving Humanity, Advancing Technology*, Atlanta, GA, USA, October 13-16 (1999)
14. LaCourse, R.J., Hudik Jr., C.F.: An eye movement communication-control system for the disabled. *IEEE Trans. Biomed. Eng.* 37(12), 1215–1220 (1990)
15. Rau, G., Schulte, E., Disselhorst-Klug, C.: From cell to movement: to what answers does EMG really contribute? *Journal of Electromyography and Kinesiology* 14, 611–617 (2004)
16. Soerberg, L.G., Knutson, M.L.: A Guide for use and interpretation of kinesiologic Electromyographic data. *Physical Therapy* 80(5), 485–498 (2000)
17. Stegeman, F.D., Blok, H.J., Hermens, J.H., Roeleveld, K.: Surface EMG models: properties and applications. *Journal of Electromyography and Kinesiology* 10, 313–326 (2000)
18. Trejo, L.J., Wheeler, K.R., Jorgensen, C.C., Rosipal, R., Clanton, S.T., Mathews, B., Hibbs, A.D., Matthews, R., Krupka, M.: Multimodal Neuroelectri interface development. *IEEE Transaction on Neural System and Rehabilitation Engineering* 11(2), 199–204 (2003)
19. Ten Kate, H.J., van der Meer, M.P.: An electro ocular switch for communication of the speechless. *Med. Progress through Technology* 10, 135–141 (1983)

An Assistive Bi-modal User Interface Integrating Multi-channel Speech Recognition and Computer Vision

Alexey Karpov, Andrey Ronzhin, and Irina Kipyatkova

St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences
SPIIRAS, 39, 14-th line, 199178, St. Petersburg, Russian Federation
{karpov, ronzhin, kipyatkova}@iias.spb.su

Abstract. In this paper, we present a bi-modal user interface aimed both for assistance to persons without hands or with physical disabilities of hands/arms, and for contactless HCI with able-bodied users as well. Human being can manipulate a virtual mouse pointer moving his/her head and verbally communicate with a computer, giving speech commands instead of computer input devices. Speech is a very useful modality to reference objects and actions on objects, whereas head pointing gesture/motion is a powerful modality to indicate spatial locations. The bi-modal interface integrates a tri-lingual system for multi-channel audio signal processing and automatic recognition of voice commands in English, French and Russian as well as a vision-based head detection/tracking system. It processes natural speech and head pointing movements in parallel and fuses both informational streams in a united multi-modal command, where each modality transmits own semantic information: head position indicates 2D head/pointer coordinates, while speech signal yields control commands. Testing of the bi-modal user interface and comparison with contact-based pointing interfaces was made by the methodology of ISO 9241-9.

Keywords: Multi-modal user interface; assistive technology; speech recognition; computer vision; cognitive experiments.

1 Introduction

Most of recent research in Human-Computer Interaction (HCI) has focused on equipping machines with means of communication that are used between human beings, such as speech, gestures, tactile interfaces. These interfaces are generally developed for ordinary people without disabilities; however, there is a lack of HCI research towards interfaces that are specifically developed for diverse groups of handicapped people. It is clear that a hearing-impaired person cannot use a speech interface, whereas a hand-disabled person cannot use a manual gesture interface and keyboard/mouse devices. Nowadays the world society pays much attention to the problems of physically and mentally handicapped persons with partial or full dysfunctions of their body parts and organs. There are several kinds of physical disabilities, manifesting themselves in impairments of speech, hearing, vision, and motion impairments such as walking or moving fingers. Many governmental programs have been launched for social and professional rehabilitation and support of disabled people.

For instance, because of a disaster, or inborn disabilities some people are unable to operate a personal computer and type by a keyboard or a mouse/touchpad due to disabilities of their hands/arms. It much restricts their interaction abilities with diverse information system and results in reducing social status. In our research, it is proposed one solution for these persons; it is a bi-modal system, which allows controlling a computer without the traditional control devices, but using: (1) head (or face) motions to control the mouse cursor through the monitor screen; (2) speech input for giving the control commands. This system relates to the class of multi-modal user interfaces and systems, which are aimed to recognize naturally occurring forms of natural language and behavior, and incorporate several recognition-based technologies (speech recognition, image analysis and computer vision, handwriting recognition etc.) [1]. Multimodal user interfaces can process two or more combined natural user input modes such as speech, touch, manual gestures, or head and body movements, in a coordinated manner with multimedia system output and allow choosing an accessible way of interaction in a concrete application for each concrete user.

The first multi-modal user interface integrating two modalities such as voice information and manual gestures for personal HCI has been proposed by R. Bolt in early 80s [2]. Probably, the first attempt to develop an assistive bi-modal interface employing head pointing and speech control for hands-free HCI oriented to ordinary and impaired people was made in late 90s [3]. There are also more recent similar assistive systems; for instance, in [4] it is proposed a vision-based uni-modal user interface for hands-free HCI based on user's nose detection and tracking by a computer vision system; in [5] it is proposed a voice-based assistive user interface (Vocal Joystick); another research in [6] presents a bi-modal user interface, based on voice input and head tracking, for some home appliances.

In the given paper, it is proposed a novel multi-modal user interface that integrates multi-channel speech recognition and computer vision technologies. Any person, who has troubles with standard computer input devices (e.g. mouse or keyboard) could manipulate mouse cursor by moving head and giving speech command instead of clicking buttons. The bi-modal interface combines modules for automatic speech recognition and head tracking in one system.

2 Architecture of the Assistive Bi-modal User Interface

The proposed bi-modal user interface has been named ICANDO (an Intelligent Computer AssistaNt for Disabled Operators) and it is intended mainly for assistance to persons without hands or with disabilities of their hands/arms. ICANDO integrates one software module for automatic recognition of multi-lingual voice commands in English, French and Russian, as well as another vision-based module for head detection and pointing. It processes natural human's speech and head motions in parallel and then fuses both informational streams in a joint multi-modal command for operating graphical user interface (GUI) of a computer. Each of the modalities transmits own semantic information: head position indicates 2D head/pointer coordinates, while speech signal yields control commands, which must be performed with an object selected by the pointer or irrespectively to its position. The multi-modal user interface has been implemented in two different versions:

1. Low-cost computer interface [7], which is available for most of potential users with any computers. It employs a standard web-camera priced under 50 \$. USB web-camera Logitech QuickCam for Notebooks Pro is used in this version. It captures both one-channel video signal in 640x480x25fps mode and one-channel audio signal with 16 KHz sampling rate and mono format with acceptable SNR via the built-in microphone.
2. Advanced multi-modal interface (Figure 1), which simultaneously processes both one-channel video signal obtained by the web-camera Logitech and multi-channel audio signal captured by an array of four professional microphones Oktava connected to the external multi-channel sound board Presonus Firepod.

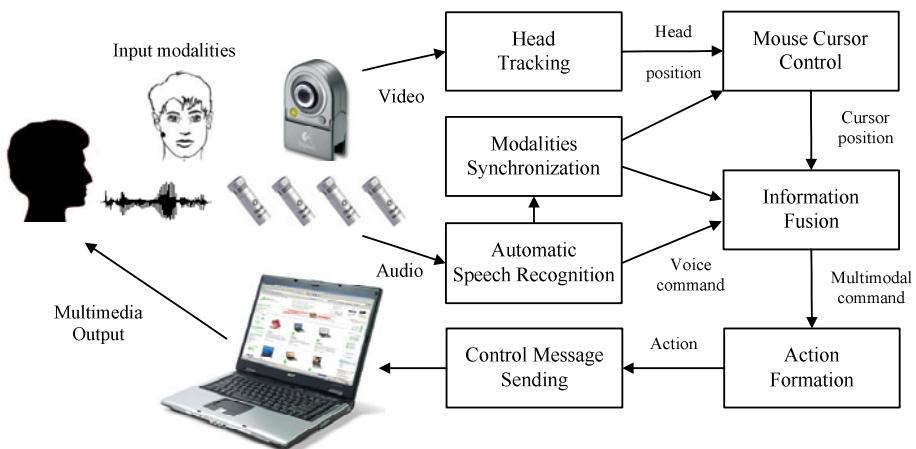


Fig. 1. Software-hardware architecture of the assistive bi-modal user interface

SIRIUS (SPIIRAS Interface for Recognition and Integral Understanding of Speech) speech recognition engine [8] is applied for automatic recognition and interpretation of input voice commands. For the speech parameterization Mel-frequency cepstral coefficients (MFCC) with 1st and 2nd derivatives are used. Modeling and recognition of phonemes and words in the system are based on the first-order continuous Hidden Markov Models (HMM) [9]. The system has been trained to understand 30 voice commands in each of 3 working languages in the speaker-dependent mode. Speaker-dependent automatic speech recognition is more adequate for the present task rather than a speaker-independent one, since it provides a lower word error rate. All the voice commands are divided into four classes according to their functional purpose: pointer manipulation commands (“Left/Click”, “Double click”, “Scroll up”, etc.); keyboard commands (“Enter”, “Escape”, etc.); GUI commands (“Start”, “Next”, “Previous”, etc.), and special system commands. However, only the pointer manipulation commands are multi-modal ones and require pointer coordinates for the information fusion.

In the advanced assistive interface additionally there is a software module for multi-channel audio signal processing. It helps to improve quality of voice activity

detection (VAD) in noisy environments, for example, when there are some talking people in the room. This module performs spatial speaker localization that is based on the general cross correlation phase transform (GCC-PHAT) method [10] and applies a microphone array in “reversed T” shape (Figure 2, left) [11] consisting of 4 cardioid microphones Oktava MK-012 (Figure 2, right) connected to a laptop via the FireWire interface of the multi-channel sound board Presonus Firepod. The construction of the microphone array is made as a rigid construction in the reversed T-shape, where 3 microphones of 4 are located linearly between the external sound board box, lying on the table, and bottom part of the laptop, whereas the fourth upper microphone locates horizontally above the laptop screen. Microphone cables are plugged into the sound board behind the laptop construction. Thus, all the microphones are maximally far from each other and such microphone’s non-linear configuration allows calculating 3D coordinates of sound sources. Estimation of correlation maximum of the mutual spectrum for all the signal pairs allows evaluating phase difference between speech signals in array channels. Further calculation of 3D speech source coordinates is made by the triangulation method. Restriction of a working zone for a user in front of the laptop screen to 0.7 m (radius of the zone) allows the system to eliminate outer acoustic noises and discard useless speech of other people improving quality of voice activity detection. The developed microphone array provides error of speech source localization less than 5°.



Fig. 2. Model of microphone array (left); one microphone of the array (right)

In the proposed bi-modal user interface, the web-camera Logitech is used jointly with the software module of computer vision based on OpenCV library in order to detect and track natural operator’s head gestures instead of hand-controlling motions. It captures raw video data in 640x480x25 fps format. At the system start, user’s head is automatically searched in video frames, employing an object detection based on Viola-Jones method [12] with Haar-like features of human’s head model [13]. It is able to find rectangular regions in a given image that likely contain human’s face. Region of interest has to be larger than 220x250 pixels in frames of 640x480 points allowing the system to find only one closest and biggest head in image, accelerating visual processing. Then the computer vision system processes optical flow for continuous tracking of five natural facial markers: tip of nose, center of upper lip,

point between eyebrows, left eye (iris) and right eye. The head tracking method applies a basic iterative Lucas-Kanade algorithm [14] for optical flow with the pyramidal implementation [15]. A mouse pointer controlled by user's nose was also proposed before, for instance in [4]; however, the set of 5 facial points improves robustness of face tracking [7].

Synchronization of two information streams (modalities) is performed by the speech recognition module, which sends messages to store pointer coordinates, calculated by the head tracking module, and for information fusion. Pointer 2D coordinates are taken at a moment of start triggering the voice activity detector, instead of the moment of completion of speech recognition process. It is connected with the problem that, when speaking, a user involuntarily slightly moves his/her head so that, at the end of the speech command recognition, the pointer may indicate another graphical object.

ICANDO uses a late-stage semantic architecture [16], and the recognition results involve partial meaning representations, which are integrated by the information fusion component controlled by the dialogue manager. Fields of a semantic frame (speech command index, X coordinate, Y coordinate, command type) are filled in with the required information, and an action corresponding to the multi-modal command is made on completion of speech recognition. ASR module operates in the real-time mode ($< 0.1xRT$); since the vocabulary of voice commands is small, there are minor delays between an utterance and fulfillment of corresponding multi-modal command. If a speech command is real multi-modal one (pointer manipulation commands only) then it is combined with stored coordinates of the pointer and a message to the mouse virtual device is sent. If a voice command is uni-modal one, coordinates are not taken into account and a message to the keyboard device is posted.

The developed assistive bi-modal system has been installed on a laptop with a four-cored processor Intel Core2Quad and a wide screen of 15". The multi-modal system takes advantage of the multi-cored architecture of PC in this case.

3 Cognitive Experiments on Human-Computer Interaction

Quantitative evaluation of the developed hand-free interface was carried out using the methodology of ISO 9241-9 [17], which is based on Fitts' law experiments [18] and related works [19]. ICANDO has been quantitatively evaluated by six volunteers, including four beginners in hands-free HCI and two developers/experts. Working with ICANDO the subjects seated at a table about 0.5 meters far from the 15" laptop's screen. Prior to the experience, subjects are shown a short demonstration of the task to be performed. Then, the subjects are allowed a short training period, and instructed to click targets as quickly as possible (in order to comply with Fitts' law hypothesis). The users were instructed to point and to select 16 ordered targets/circles, with a circular layout (Figure 3, left) [20], so pointing movements must be carried out in different directions. When selection occurs, the former target is shadowed and the next target is displayed. Figure 3 (right) shows one sample of real trajectory of mouse cursor movement at hands-free HCI with head pointing.

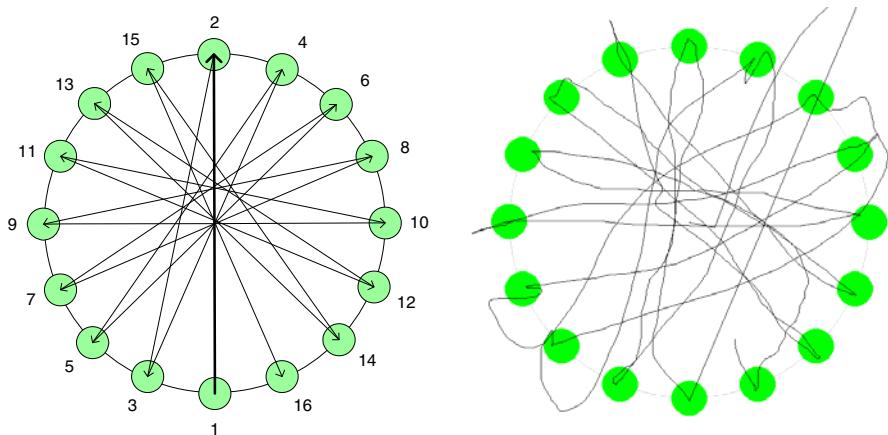


Fig. 3. Layout of round targets on screen for ISO 9241-9 based experiments (left); Example of trajectories of cursor movement at hands-free HCI with head pointing (right)

The experiments with several pairs of targets width/distance, corresponding to different indexes of difficulty were carried out by each subject. The index of difficulty ID of the task, measured in bit by $ID = \log_2(D/W + 1)$, where D is the distance between targets and W is the target's width. In the experiments ID varied in the range from 1.32 till 4.4 (10 different values). However, a location where selection occurs influences both on effective distance and effective width. The effective index of difficulty is $IDe = \log_2(De/We + 1)$, where De is the effective distance between the first and last points of a pointer trajectory; We is the effective target width: $We = 4.133\sigma$, where σ is the standard deviation of the coordinates of the point of selection, projected on the axis between the centers of the origin and destination targets [21].

Fitts' law states that the movement time MT between two targets is a linear function of ID of the task related to the targets' characteristics. Figure 6 shows the movement time MT versus the effective index of difficulty IDe for ICANDO system. For each trial, the inter-target movement time is defined and measured as the time between two successive selection events, selection occurs counted both inside and outside targets (selection error). Figure 4 plots dependences between MT and IDe for two groups of users (beginners and experts). These plots show that trained users capable to perform the given task much faster than novices; however, the same is true for any contact-based pointing devices as well.

Throughput TP is the linear ratio between IDe and MT measured in bits per second [22]. Mean TP allows comparison between performances of different pointing interfaces. Standard contact-based pointing devices such as a mouse, touchpad, trackball, joystick and 17" touchscreen were also evaluated by the ISO 9241-9 standard in order to compare their performances with those of the proposed contactless interface. Table 1 shows averaged values of movement time MT and effective throughput TPe , which is a tradeoff between pointing speed and target selection precision.

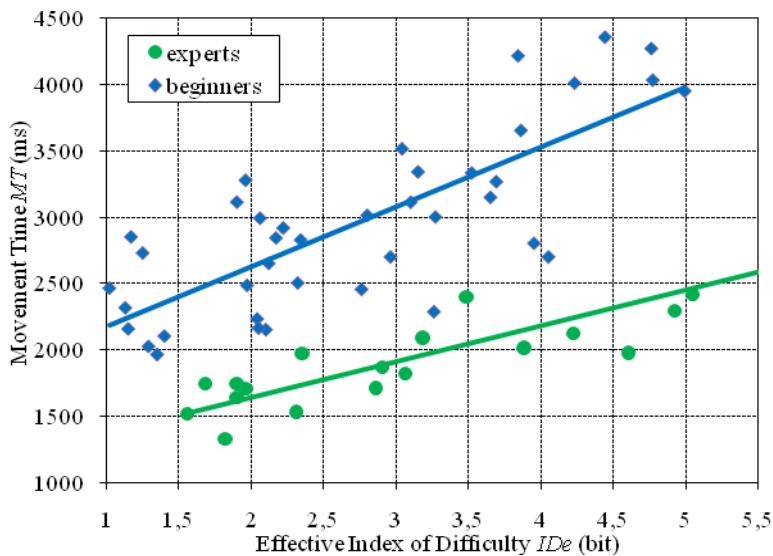


Fig. 4. Dependence of Movement Time MT on Effective Index of Difficulty IDe for two groups of users

The best TPe results have been obtained with the contact-based touch screen and the optical mouse device, taking into account that a touch screen is not so precise for small W . Performance results obtained with ICANDO interface, as well as the joystick and trackball devices are rather similar. However, the main advantage of the developed bi-modal interface is that it provides natural hand-free way of human-computer interaction.

Table 1. Quantitative comparison of performances of pointing interfaces using cognitive Fitts' law experiments

Interface of HCI	MT , seconds	Selection error, %	TP , bit/second
Joystick	2.01	7.00	1.54
Trackball	1.03	3.83	3.51
Touchpad 3"	0.85	4.50	3.72
Optical mouse	0.49	3.17	6.65
Touchscreen 17"	0.50	6.17	7.85
Head pointing + voice	1.98	7.33	1.59

Moreover, one experienced human being applied the basic bi-modal assistive system for the Internet surfing task in several HCI sessions during one day. A statistical analysis of the system's log has shown that about 700 meaningful voice control commands were issued excluding some out-of-vocabulary words and fails of

VAD. However, some speech commands were more frequent than others, and some other commands were not used at all. Figure 5 presents a relative distribution of the issued voice commands in this cognitive experiment.

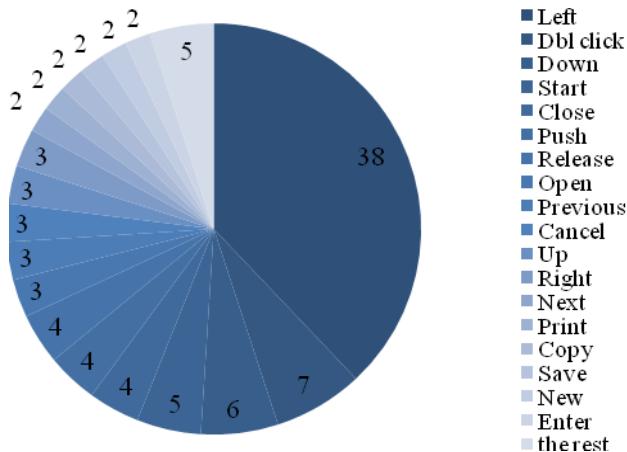


Fig. 5. Relative distribution of voice commands usage at experiments on hands-free HCI

It was predictably that the most important and frequent speech command is “Left” (more than 1/3 of all cases), because it replaces click of the left mouse button that is used very often (for instance, when typing letters in fields of some windows and forms). Internet surfing task sometimes requires to enter text data (for instance, URLs) and user can do it by ICANDO and virtual On-Screen keyboard, confirming pointed letters by the command “Left” (alternatively, Dasher [23] data entry contactless interface can be applied as well). All the other voice commands are distributed among the rest space being 7% at the maximum. Totally above 64% of the speech commands were given in the multi-modal way (simultaneously with the head pointing) and the rest of the commands were uni-modal speech-only commands.

4 Conclusion

The developed bi-modal user interface integrating automatic speech recognition and computer vision technologies was tested by cognitive experiments both by able-bodied human beings and a person with a severe disability (specifically the user has no hands). Video demonstrations of this assistive system are available in WWW: [24] and [25]. In order to quantitatively evaluate the hands-free pointing interface, we used the methodology of ISO 9241-9, which is based on the Fitts' law experiments. Comparisons of the proposed hands-free interface with the contact-based pointing devices (mouse, touchpad, trackball, touch screen and joystick) were made in terms of the effective index of difficulty, movement time and throughput parameters. The best results were shown by the contact-based touch screen and mouse devices, but the bi-modal interface has outperformed the joystick device.

The obtained performance of hand-free HCI is acceptable, since the developed bi-modal interface is intended mainly for human beings with severe motor-disabilities. ICANDO allows supporting equal participation and socio-economic integration of people with disabilities in the information society and improving their independence from other people. However, it can be helpful for ordinary users too for hands-free human-computer interaction in diverse applications, where hands of a human being are busy, for instance when driving or cooking.

Acknowledgements. This research is partially supported by the Federal Targeted Program “Scientific and Scientific-Pedagogical Personnel of the Innovative Russia in 2009-2013” (contracts No. 2360 and No. 2579), by the Grant of the President of Russian Federation (project No. MK-64898.2010.8), and by the Russian Foundation for Basic Research (project No. 09-07-91220-CT-a).

References

1. Oviatt, S.: Multimodal interfaces. In: Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications, pp. 286–304. Lawrence Erlbaum Assoc., Mahwah (2003)
2. Bolt, R.A.: Put-that-there: Voice and gesture at the graphics interface. Computer Graphics 14(3), 262–270 (1980)
3. Malkewitz, R.: Head Pointing and Speech Control as a Hands-Free Interface to Desktop Computing. In: 3rd International ACM Conference on Assistive Technologies ASSETS 1998, Marina del Rey, CA, USA, pp. 182–188. ACM Press, New York (1998)
4. Gorodnichy, D., Roth, G.: Nouse 'Use your nose as a mouse' perceptual vision technology for hands-free games and interfaces. Image and Vision Computing 22(12), 931–942 (2004)
5. Harada, S., Landay, J.A., Malkin, J., Li, X., Bilmes, J.A.: The Vocal Joystick: Evaluation of voice-based cursor control techniques. In: 8th International ACM SIGACCESS Conference on Computers & Accessibility ASSETS 2006, Portland, USA, pp. 197–204. ACM Press, New York (2006)
6. Eiichi, I.: Multi-modal interface with voice and head tracking for multiple home appliances. In: 8th IFIP International Conference on Human-Computer Interaction INTERACT 2001, Tokyo, Japan, pp. 727–728 (2001)
7. Karpov, A., Ronzhin, A.: ICANDO: Low Cost Multimodal Interface for Hand Disabled People. Journal on Multimodal User Interfaces 1(2), 21–29 (2007)
8. Ronzhin, A., Karpov, A.: Russian Voice Interface. Pattern Recognition and Image Analysis: Advances in Mathematical Theory and Applications 17(2), 321–336 (2007)
9. Rabiner, L., Juang, B.: Speech Recognition. In: Benesty, J., et al. (eds.) Springer Handbook of Speech Processing. Springer, New York (2008)
10. Krim, H., Viberg, M.: Two decades of array signal processing research: the parametric approach. Signal Processing Magazine 13(4), 67–94 (1996)
11. Ronzhin, A., Karpov, A., Kipyatkov, I., Železný, M.: Client and Speech Detection System for Intelligent Infokiosk. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2010. LNCS, vol. 6231, pp. 560–567. Springer, Heidelberg (2010)
12. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: IEEE International Conference on Computer Vision and Pattern Recognition Conference CVPR 2001, Kauai, HI, USA (2001)

13. Lienhart, R., Maydt, J.: An Extended Set of Haar-like Features for Rapid Object Detection. In: IEEE International Conference on Image Processing ICIP 2002, Rochester, New York, USA, pp. 900–903 (2002)
14. Lucas, B.D., Kanade, T.: An Iterative Image Registration Technique with an Application to Stereo Vision. In: 7th International Joint Conference on Artificial intelligence IJCAI, Vancouver, Canada, pp. 674–679 (1981)
15. Bouguet, J.-Y.: Pyramidal Implementation of the Lucas-Kanade Feature Tracker Description of the algorithm. Intel Corporation Microprocessor Research Labs, Report, New York, USA (2000)
16. Karpov, A., Ronzhin, A., Cadiou, A.: A Multi-Modal System ICANDO: Intellectual Computer AssistaNt for the Disabled Operators. In: INTERSPEECH International Conference, Pittsburgh, PA, USA, pp. 1998–2001. ISCA Association (2006)
17. ISO 9241-9:2000(E) Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs), Part 9: Requirements for Non-Keyboard Input Devices, International Standards Organization (2000)
18. Soukoreff, R.W., MacKenzie, I.S.: Towards a standard for pointing device evaluation, perspectives on 27 years of Fitts' law research in HCI. International Journal of Human Computer Studies 61(6), 751–789 (2004)
19. Zhang, X., MacKenzie, I.S.: Evaluating Eye Tracking with ISO 9241 - Part 9. In: Jacko, J.A. (ed.) HCI 2007. LNCS, vol. 4552, pp. 779–788. Springer, Heidelberg (2007)
20. Carbini, S., Viallet, J.E.: Evaluation of contactless multimodal pointing devices. In: 2nd IASTED International Conference on Human-Computer Interaction, Chamonix, France, pp. 226–231 (2006)
21. De Silva, G.C., Lyons, M.J., Kawato, S., Tetsutani, N.: Human Factors Evaluation of a Vision-Based Facial Gesture Interface. In: International Workshop on Computer Vision and Pattern Recognition for Computer Human Interaction, Madison, USA (2003)
22. Wilson, A., Cutrell, E.: FlowMouse: A computer vision-based pointing and gesture input device. In: Costabile, M.F., Paternó, F. (eds.) INTERACT 2005. LNCS, vol. 3585, pp. 565–578. Springer, Heidelberg (2005)
23. Ward, D., Blackwell, A., MacKay, D.: Dasher: A data entry interface using continuous gestures and language models. In: ACM Symposium on User Interface Software and Technology UIST 2000, pp. 129–137. ACM Press, New York (2000)
24. SPIIRAS Speech and Multimodal Interfaces Web-site, TV demonstration,
<http://www.spiiras.nw.ru/speech/demo/ort.avi>
25. SPIIRAS Speech and Multimodal Interfaces Web-site, demonstration 2,
http://www.spiiras.nw.ru/speech/demo/demo_new.avi

A Method of Multiple Odors Detection and Recognition

Dong-Kyu Kim¹, Yong-Wan Roh², and Kwang-Seok Hong¹

¹ Human Computer Interaction Laboratory, School of Information & Communication Engineering, Sungkyunkwan University, 440-746, 300 CheonCheon-Dong, Jangan-Gu, Suwon, Gyeonggi-Do, Korea,

² Information and Communication Engineering, Seoil University, 131-702, #49-3 Myeonmok-dong, Seoildaehak-gil-22, Jungnang-gu, Seoul, Korea,
`{kdgyu13, elec1004}@skku.edu, kshong@skku.ac.kr`

Abstract. In this paper, we propose a method to detect and recognize multiple odors, and implement a multiple odor recognition system. Multiple odor recognition technology has not yet been developed, since existing odor recognition techniques which have been researched and developed by components analysis and pattern recognition techniques only deal with single odors at a time. Multiple odors represent a dynamic odor change from no odor to a single odor and multiple odors, which is the most common situation in a real-world environment. Therefore, it is necessary to sense and recognize techniques for dynamic odor changes. To recognize multiple odors, the proposed method must include odor data acquisition using a smell sensor array, odor detection using entropy, feature extraction using Principal Component Analysis, recognition candidate selection using Tree Search, and recognition using Euclidean Distance. To verify the validity of this study, a performance evaluation was conducted using a 132 odor database. As a result, the odor detection rate is approximately 95.83% and the odor recognition rate is approximately 88.97%.

Keywords: Dynamic Odor Change, Multiple Odors, Odor Detection and recognition.

1 Introduction

Smelling is an activity that uses olfactory information, which takes place when an odorant transferred inside our nose stimulates the chemical sensory receptors. Among the five senses, olfactory information provides the most powerful memory and association mechanism and intensification effects on other senses [1]. Because of this phenomenon, products and marketing use odors that may stimulate consumer emotions. Therefore, the sense of smell is expected to be positioning itself as the essential sensory element in the future emotion industry. In order to make progress in using odors for commercial purposes, the development of odor recognition technology must take precedence. However, smell recognition technology in the past had been developed as a technology to recognize only one odor, so the technology to recognize multiple odors has not yet been developed. This is unfortunate, because in

the real-world environment multiple odors are the norm. For example, it is easy to encounter a situation in which the scent of flowers, wine, food, and tea are mixed together, each odor occurring in various concentrations.

One odor has a simple pattern where the relevant stimulation takes place and gets eliminated [2-4]. However, with multiple odors more than two odors may occur at the same time, or the odors may occur sequentially, that is, the second one follows the first one after a little while. On the other hand, both odors may be eliminated simultaneously, as well as the first one and second one may be eliminated sequentially. Therefore, in a odor detection stage of odor recognition, an odor may have two inflection points (input and elimination of the odor) at the most, which makes it possible for us to detect the smell. But with multiple odors, the number of inflection points may be four at the least (the input point of the first one, the input point of the second one, the elimination point of the first one and the elimination point of the second one), so it is difficult to determine multiple odors. In the odor recognition stages, one odor may have two patterns (the input of one odor and the elimination of one odor). But in the case of multiple odors, there may be more than six patterns (the input of one odor, the simultaneous input of multiple odors, the input of one odor on top of another odor, the elimination of one odor from multiple odors, the simultaneous elimination of multiple odors, or the elimination of one odor from the different odor). Due to such reasons, there is difficulty in recognizing multiple odors which show a dynamic change in odors.

For this reason, this study was designed to suggest the following four items, in order to detect and recognize multiple odors showing dynamic changes in odors. 1) Suggest an odor detection technology in which entropy is used. 2) Suggest a method with which we can recognize multiple odors. 3) Suggest a multiple odor tree structure. 4) Suggest a tree search based on the tree structure.

In order to recognize multiple odors, in this study, the system of odor recognition was organized, which includes the acquisition of odor data, odor detection, feature extraction and odor recognition. The acquisition of odor data is the stage where the odor is converted to the odor data by using the sensor array. Odor detection is the stage where the inflection point (the time when the input and elimination of odor is carried out) of an odor is detected using the entropy derived from odor data. And feature extraction is the stage where the feature vector of the odor pattern is extracted using Principal Component Analysis (PCA). Finally, odor recognition is the stage where the recognition candidates are selected by using a tree search, and the odor of which the Euclidean distance of the feature vector is the closest is judged as the recognition result. Based on such stages mentioned above, the experiment was carried out. The composition of this paper is as follows. In Chapter 2, the relevant studies are explained, and in Chapter 3, the system of odor recognition is introduced. In Chapter 4, the experimental content that is aimed at evaluating the suggested system is described. Finally, in Chapter 5, the conclusion is drawn.

2 Related Studies

The odor sensor employed by one of the current odor recognition systems remains at the technological level capable of responding to a specific gas, obtaining odor data by

constructing an array of odor sensors of different kinds in order to recognize an odor. It recognizes odors using the processes of odor detection, feature extraction, and odor classification from the acquired odor data. However, it has constraints in detecting and recognizing multiple odors that show dynamic changes.

The existing odor-recognition technology research uses odor sensor arrays, chambers, micro air pumps, and other equipment to obtain odor data. An odor was sucked in by turning on a micro pump after locating an odor inlet at the entrance of a bottle containing the source of the odor. It is a structure of obtaining odor data by filling a chamber with an odor. Then, the odor that filled the chamber is eliminated by attaching the odor-suction inlet to a location away from the entrance of the bottle containing the odor source [5]. The odor data acquisition method as such is useful as a method for obtaining and recognizing a single odor, but it can't be used to detect dynamically-changing multiple odors.

There are several types of sensors that are commonly used to detect odors, such as a Metal Oxide Sensor (MOS), Conducting Polymer (CP), Quartz Crystal Microbalance (QCM), and Surface Acoustic Wave (SAW). In recent studies, the MOS is the most popular sensor, because it is small and has high sensing performance. The entropy technique, which is an application of the conventional odor detection methods, is able to detect the input of an odor, but it does not suggest a way of detecting odor inputs and deodorization against multiple dynamically-changing odors. The reported feature-extraction and pattern-recognition algorithms used in the process of odor-recognition are PCA, Linear Discriminant Analysis (LDA), Neural Networks (NN), or k-Nearest Neighbor (k-NN) [6-12]. Though such feature-extraction and pattern-recognition algorithms are useful for odor recognition, as the number of the targets of recognition increases, the recognition performance may decline. In order to increase recognition performance, new methods such as recognition candidate selection capable of discerning multiple odors, a single odor, and an odorless state are necessary.

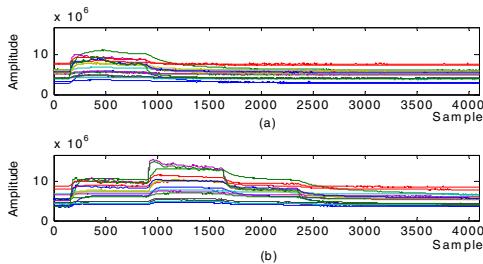


Fig. 1. Single odor and multiple odor patterns: (a) single odor pattern, (b) multiple odor pattern

As it turned out, additional technologies are required for the processes of odor data acquisition, odor detection, and odor recognition, in order to detect and recognize multiple odors showing dynamic changes. That is, there are differences between a single-odor pattern and multiple-odor patterns acquired from an odor data acquisition device. Figure 1 indicates a single-odor pattern and multiple-odor patterns. The (a) in Figure 1 is a case of a single-odor pattern eliminated after the odor is input, and (b) is a case of multiple odor patterns in which odors are sequentially eliminated one by one under the condition of being multiple odors by iteratively inputting odor after odor.

3 Multiple Odor Recognition System

As shown in Figure 2, the total system consists of: acquisition of odor data, detection of odor, extraction of characteristic, and recognition of odor. Odor data is acquired through a sensor array converting and saving the odor into odor data. The odor is detected using entropy from measured odor signals, and detects the inflection points of odor (odor input and removal point). Characteristics are extracted by using PCA to extract the characteristic vector of an odor pattern. The odor is recognized using a Tree Search to select the recognition candidate, and by a Euclidian distance value.

The odor extraction method monitors the abrupt change in the sensor amplitude during odor input and removal, and extracts the odor from the change in information using entropy. The feature extraction technique using PCA for odor recognition searches for the axes of features from input data of high dimensions, which becomes the main element of low dimension, and then, extracts features by a projection based on these main axes [13-16]. Multiple odors have a tree-structure that divides possibilities into odorless, single odor, and multiple odors depending on the changes of odor input and removal conditions. The tree search method uses the amplitude information in the initial point of the odor pattern, amplitude information in the final point, and information from the previous odor pattern to determine how the present odor pattern has changed. It also reduces the number of patterns on the subject for recognition. It searches the changes from odorless to single odor or multiple odors, the change from single odor to odorless or multiple odors, and the change from multiple odors to odorless or single odor. It determines the odor that has is closest between the characteristic vector of input odor and the one of recognition candidate as the result of recognition.

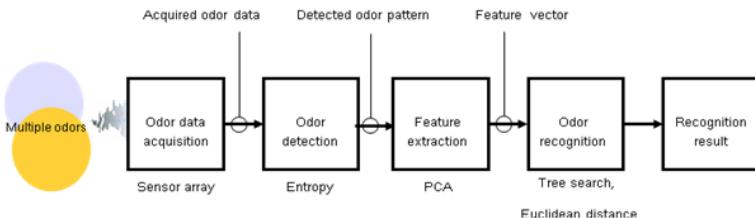


Fig. 2. Odor recognition system block diagram

3.1 Odor Data Acquisition

In this paper, odor data became obtainable using several different kinds of sensors for dynamically-changing odors in real time. In order to obtain odor data in regard to dynamically-changing odors, a structure was designed and manufactured to allow the surrounding air to pass through it in real time by attaching fans in front of and behind a chamber. This structure of inhaled and discharged the surrounding air containing chemicals that constitute a specific odor into and out of the chamber. The structural drawing of such odor-data acquisition is displayed in Figure 3.

Odor data was acquired by putting together 16 different types of MOS as an array in this study. Sensor circuits consisted of a partial pressure resistance circuit that converts the resistance change of the sensor for odor into voltage. For A/D conversion, MCU of C8051F350 was used, and to send out the acquired data to the recognition server, a Bluetooth communication module was included. The A/D conversion speed is 13.5 times per second, and it was sent out with a speed of 115.2Kbps using RS232. The manufactured size of the odor data acquisition device had a width of 45mm, length of 160mm, and height of 90mm.

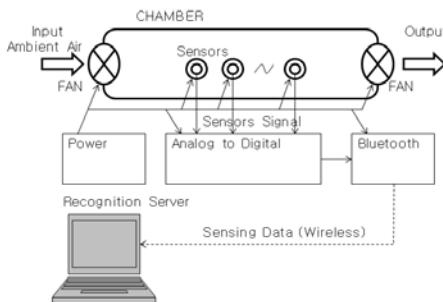


Fig. 3. The structural drawing of such odor-data acquisition

3.2 Odor Detection Using Entropy

Odor was detected using entropy, determining the inflection point (odor input and removal point). Odor was detected by monitoring the abrupt change in the amplitude of a sensor during odor input and removal. This is detected by monitoring the entropy. To detect the inflection point of multiple odors using entropy, the sampling frequency per channel was 13.5Hz. 1 Frame had 32 Samples, an overlap of 16 Samples, a calculated entropy value from each Frame. This information was used to figure out the number of sensors that had an entropy value exceeding 1, the condition with more than 7 sensors that have entropy exceeding 1, and continuous condition of more than 4 Frames for entropy and sensor reaction to reduce wrongful extraction. By adding continuous conditions of frames, the abrupt rise in entropy values due to the noise between the actual inflection point and the inflection point was eliminated, and thus could prevent wrongful extraction. A total of 256 samples, from the first 32 samples to the 224 samples after the starting point, were selected as the domain for characteristic extraction, based on the extracted inflection point. Figure 4 shows the inflection point extraction procedure using entropy. For odor detection, the sensor response to the input odor was applied to the entropy information entropy, which was defined by C. Shannon, for determination of a change point of sensor measurement values on input-odor. Entropy in the frame unit for an individual sensor can be calculated by the following Equation (1).

$$E(n) = - \sum_{i=0}^{L-1} p(i) \times \log_2 p(i) \quad (1)$$

Where $p(l)$ is the probability of a certain range l occurring in one frame. The occurrence rate $c(l)$ of the range l in one frame is divided by the number of samples s in one frame as shown in Equation (2).

$$p(l) = \frac{c(l)}{s} \quad (2)$$

Where l can be shown as $0 \leq l < L - 1$, meaning that the response range of a sensor is divided into numbers of L .

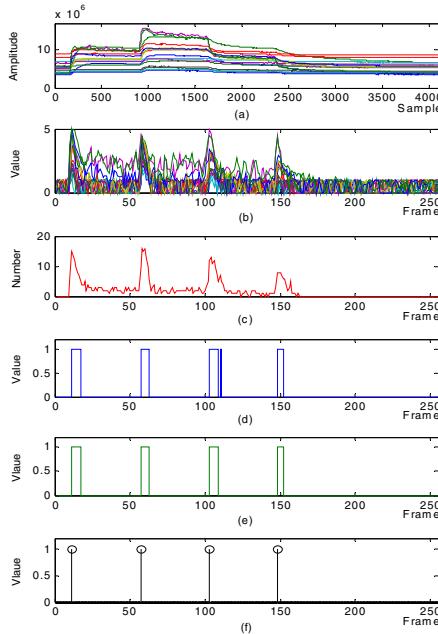


Fig. 4. Inflection point extraction procedure using entropy: (a) multiple odors pattern, (b) entropy value for each sensor signal, (c) number of sensors with entropy exceeding Threshold (over 1), (d) response frame with more than 7 sensors with entropy exceeding Threshold (over 1), (e) result frame with number of continuous frames with more than 4 Frames, (f) inflection point of multiple odors

3.3 Recognition Candidate Selection Using Tree Search

In order to look into the changes in the state of odors, a tree search was employed to search for changes in the state of multiple odors. Tree Search determines how the present odor pattern has changed by using the odor pattern's starting point amplitude information, endpoint amplitude information, and information on the previous odor pattern condition (previous recognition result). When an odor is input, the amplitude increases, and when odor is removed, the amplitude decreases. Thus, when the starting point amplitude is smaller than that of the end point, it can be determined as

an odor input, and when the starting point amplitude is bigger than that of the end point, it can be determined as an odor removed. Multiple odors, a single odor, and the odorless state are determined according to the result of conversion from the previous odor pattern to the present condition. The condition for determining a single odor is when one odor is input in an odorless condition, and when one of the two odors is removed. The condition for determining multiple odors is when multiple odors are input in an odorless condition, and when another odor is input in a single odor. The condition for determining an odorless state is when during the initial odorless condition, one odor is removed from a single odor condition, and when multiple odors are removed from a multiple odors condition. Thus, the conditions such as when odorless becomes a single odor or multiple odors, a single odor is changed to odorless or multiple odors, and when multiple odors are changed to odorless or single odor are searched. Tree Search consequently decreases recognition candidates. Figure 5 shows the tree structure of multiple odors.

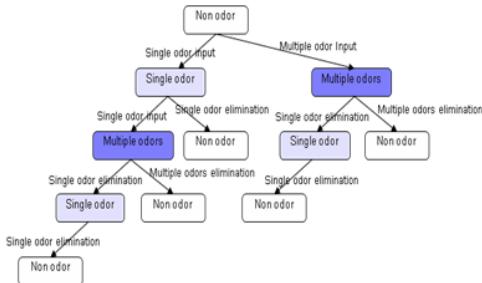


Fig. 5. Tree structure of multiple odors

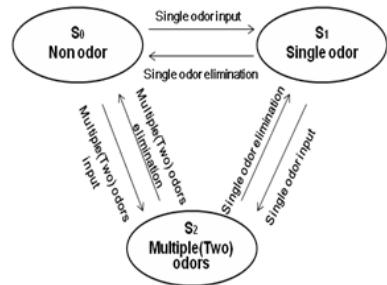


Fig. 6. Changes in multiple odor conditions

The Tree Search method based on the suggested tree structure uses the condition information of input and removed odors. The changes in multiple odor conditions as shown in Figure 6, S_0 condition is odorless, S_1 condition is a single odor, and S_2 condition is two odors. Odor conditions can change from S_0 to S_1 or S_2 , or from S_1 to S_0 or S_2 , or from S_2 to S_0 or S_1 .

4 Experiment and Results

In this paper, experiments for detecting and recognizing odorlessness, single odors, and two odors were performed. The objective of the experiments was to collect multiple odor databases against various cases in which multiple odors occur and then were eliminated. And, experiments for detecting and recognizing odors using collected databases were performed. For that reason, 11 scenarios containing various occurrences and eliminations of multiple odors were selected whereby multiple-odor databases were collected. Table 1 indicates the multiple odors generation and elimination scenarios.

Table 1. The multiple odors generation and elimination scenarios

No.	The multiple odors generation and elimination scenarios	No. of inflection point
1	A odor generation → B odor generation → B odor elimination → A odor elimination	4
2	A odor generation → B odor generation → A odor elimination → B odor elimination	4
3	B odor generation → A odor generation → A odor elimination → B odor elimination	4
4	B odor generation → A odor generation → B odor elimination → A odor elimination	4
5	A and B odors generation → A odor elimination → B odor elimination	3
6	A and B odors generation → B odor elimination → A odor elimination	3
7	A odor generation → B odor generation → A and B odors elimination	3
8	B odor generation → A odor generation → A and B odors elimination	3
9	A and B odors generation → A and B odors elimination	2
10	A odor generation → A odor elimination	2
11	B odor generation → B odor elimination	2

The reason why odor-database collection has been performed by applying experimental environments, experimental conditions, experimental odor sources, etc. to each experiment in existing research cases attend the lack of internationally-standardized odor databases [2-12].

In this experiment, flower odor sources generating odors safely and evenly were selected. As the experimental flower odor sources, 4 kinds of flower odor oils (100% pure and natural essential oil, professional grade, EuroAroma®) contained in a small 5 ml bottle were used.

For evaluating the efficiency of the suggested system, a total of 4 kinds of floral fragrance oils (Number 1 fragrance = LAVENDER, number 2 fragrance = HYSSOP, number 3 fragrance = GERANIUM, number 4 fragrance = ROSEMARY) were used. Fragrances 1 and 2, fragrances 1 and 3, fragrances 1 and 4, fragrances 2 and 3, fragrances 2 and 4, and fragrances 3 and 4 were crossed for database collection.

The entire database was created by collecting 132 different combinations of scents, as a result of doubling the multiplication of 6 kinds of odors by 11 kinds of scenarios. The collection procedure of odor database collection is, if the first item of Table 1 is taken as an example, performed in the following sequence: 1) Start saving, 2) Inputting odor A (at about 10 sec), 3) Inputting odor B (at about 70 sec), 4) Eliminating odor B (at about 130 sec), 5) Eliminating odor A (at about 190 sec), 6) Complete saving (at about 304 sec), 7) Ventilating.

Odor detection and recognition experiments using multiple odor databases were performed. Odor detection was represented by detection rate as an index of the performance evaluation of the number in which the inflection point was accurately detected. Odor recognition was represented by recognition rate as an index of performance evaluation of the number in which inputted odors were accurately recognized.

The 132 database items create 408 inflection points that present the changes in multiple odors, single odor, and odorless conditions. The number of subjects for extraction and recognition are the same as their corresponding inflection points, because odor detection is based on inflection points.

The odor detection experiment took place using entropy from the multiple collected odors database to recognize multiple odors collected. The total number of

inflection points of the detection subjects was 408, and the result of odor detection experiment was 391. The detection rate was about 95.83%. Table 2 shows odor recognition experiment results. The results of odor recognition experiment showed a recognition rate of about 88.97%. However, the result of the odor recognition experiment that excluded the wrongfully detected results in the detection stage showed a recognition rate of about 92.84%.

From the experimental results of odor recognition, the reason that the recognition rate of an odor was lower than that of multiple odors is that the odor sources used in the experiment were 100% flower odor oils. Although flower odor oils were used for the safety of the experiment, the flower odor oils are vegetable-based, showing similar intensity and characteristics. However, in the case of mixing two flower odors, the recognition rate was relatively high due to the changes in the intensity and characteristics of the odors. The reason that the recognition rate of an odorless state was not 100% was due to misdetection. The recognition rate of an odorless state excluding misdetection was 100%.

Table 2. Odor recognition study results

Item	Items in detail	No. of recognition subject	No. of recognition	Recognition rate (%)
Multiple odors (mixed)	Fragrance 1 and 2 (mixed)	18	16	88.89
	Fragrance 1 and 3 (mixed)	18	17	94.44
	Fragrance 1 and 4 (mixed)	18	18	100
	Fragrance 2 and 3 (mixed)	18	15	83.33
	Fragrance 2 and 4 (mixed)	18	15	83.33
	Fragrance 3 and 4 (mixed)	18	16	88.89
Single odor	Fragrance 1	42	37	88.10
	Fragrance 2	42	28	66.67
	Fragrance 3	42	37	88.10
	Fragrance 4	42	34	80.95
Odorless	Odorless	132	130	98.48
	Total	408	363	88.97

5 Conclusion

This study was about detecting and recognizing multiple odors. This study suggested and evaluated the performance of the following four subjects: 1) odor detection methods using entropy, 2) multiple odor recognition method, 3) multiple odor tree structure, and 4) tree search based on tree structure.

In this paper, an odor-recognition system composed of the acquired odor data using odor sensor arrays, odor detection using entropy, feature extraction using PCA, selection of recognition candidates using a tree search, and Euclid distance was established for recognizing multiple odors. In order to verify the validity of the proposed multiple odor-recognition technology, 132 odor databases were collected, and performance evaluation experiments using the collected databases were performed. In the case of odor detection, the detection rate was roughly 95.83%, and the recognition rate in the case of odor recognition was about 88.97%.

This suggested technology is thought to soar past existing technologies which only recognize single odors, since it is good enough to detect and recognize multiple odors. When applying this technology to various applications, the type and number of sensors, and critical values in various conditions should be properly adjusted.

Acknowledgment. This research was supported by MKE, Korea under ITRC NIPA-2010-(C1090-1021-0008) (NTIS-2010-(1415109527)) and the National Research Foundation of Korea (NRF) grant funded by the Korean government (MEST) (No. 2010-0017123).

References

1. Kim, J.-D., Kim, D.-J., Han, D.-W.: A Proposal Representation, Digital Coding and Clustering of Odor Information. In: Computational Intelligence and Security, 2006 International Conference, vol. 1, pp. 872–877 (2006)
2. Takano, M., Fujiwara, Y., Sugimoto, I., Mitachi, S.: Real-time sensing of roses' aroma using an odor sensor of quartz crystal resonators. IEICE Electronics Express 4(1) (2007)
3. Lozano, J., Santos, J.P., Aleixandre, M., Sayago, I., Gutiérrez, J., Horrillo, M.C.: Identification of Typical Wine Aromas by Means of an Electronic Nose. IEEE Sensors Journal 6(1), 173–177 (2006)
4. Zhang, Q., Xie, C., Zhang, S., Wang, A., Zhua, B., Wang, L., Yang, Z.: Identification and pattern recognition analysis of Chinese liquors by doped nano ZnO gas sensor array. Sens. Actuators B 110, 370–376 (2005)
5. Zhang, S., Xie, C., Zeng, D., Zhang, Q., Li, H., Bi, Z.: A feature extractio method and a sampling system for fast recognition of flammable liquids with a portable E-nose. Sensors and Actuators B 124, 437–443 (2007)
6. Seiko, M., Katsunori, S., Masashi, K., Iwao, S.: Odor sensing in natural environment (2)-Application to the rose aroma sensing in an outside garden. Nippon Kagakkai Koen Yokoshu 83(1) (2003)
7. Fukai, S., Abe, Y.: Discrimination of lily fragrance by use of an electronic nose. In: ISHS Acta Horticulturae 572: XX International Eucarpia Symposium, Section Ornamentals, Strategies for New Ornamentals - Part II (2001)
8. Cheon, B.-G., Roh, Y.-W., Kim, D.-J., Hong, K.-S.: An implementation of floral scent recognition system using ICA combined with correlation coefficients. In: Adali, T., Jutten, C., Romano, J.M.T., Barros, A.K. (eds.) ICA 2009. LNCS, vol. 5441, pp. 654–661. Springer, Heidelberg (2009)
9. Roh, Y.-W., Kim, D.-K., Hong, K.-S.: A Method of Optimal Sensor Decision for Odor Recognition. In: ACCS 2009 (2009)
10. Roh, Y.-W., Kim, D.-K., Cheon, B.-G., Hong, K.-S.: A Study on User Identification using Breath Odor. In: ACCS 2009 (2009)
11. Jolliffe, I.T.: Principal Component Analysis. Springer, New York (1986)
12. Martinez, A.M., Kak, A.C.: PCA versus LDA. IEEE Trans. Pattern Analysis and Machine Intelligence 23(2) (2001)
13. Smith, L.I.: A tutorial on Principal Components Analysis (February 26, 2002)
14. Scott, S.M., James, D., Ali, Z.: Data analysis for electronic nose systems. Microchim Acta 156, 183–207 (2007)

Report on a Preliminary Study Using Breath Control and a Virtual Jogging Scenario as Biofeedback for Resilience Training

Jacquelyn Ford Morie, Eric Chance, and J. Galen Buckwalter

University of Southern California
USC Institute for Creative Technologies, 12015 Waterfront Drive,
Playa Vista, CA 90094-2536
{morie, chance, jgbuckwalter}@ict.usc.edu

Abstract. Alternative methods of treating psychological stress are needed to treat some veterans of recent military conflicts. The use of virtual world technologies is one possible platform for treatment that is being explored by the “Coming Home” project at the University of Southern California’s Institute for Creative Technologies (ICT). One of the novel ways ICT is attempting to mitigate stress via virtual worlds is with a virtual jogging scenario, where the movement of an avatar is controlled via rhythmic breathing into a standard microphone. We present results from a preliminary study of 27 participants that measured the mood and arousal effects produced by engaging in this scenario.

Keywords: Breathing, Virtual World, Second Life, stress relief.

1 Introduction

With two million troops deployed to the recent conflicts of OEF/OIF and Afghanistan, it is not surprising that there has been an overall increase in the number of military personnel reporting varied health issues arising from psychological stresses encountered during their service. These issues range from confirmed diagnoses of Post Traumatic Stress Disorder (PTSD) to more general problems such as anxiety, sleep disorders, substance abuse, marital problems, aggressive behavior and other difficulties [1]. One in five returned service members are estimated to be suffering from PTSD or some form of stress that affects their mental health [2]. At the USC Institute for Creative Technologies (ICT), we have been working on a project called “Coming Home” that researches and develops novel techniques to bring stress relief activities to our returned troops. Our goal is not to replace current treatments, but to find alternative and innovative methods that can supplement standard treatments in ways that overcome known challenges to getting help, such as perceived social stigma, disability, and limited access to health care facilities.

The military is already taking steps to combat the first of these three barriers, the stigma toward treatment. The Defense Center for Excellence in Psychological Health and Traumatic Brain Injuries (DCoE/PH) is helping to address this by providing commanders both an awareness of the negative impact such stigma has on continued

troop health, and new tools through projects such as Real Warriors (<http://www.realwarriors.net/materials>) to help mitigate its persistence.

It is more problematic to address the other two barriers –disability and limited access– to seeking treatment. Soldiers with physical injuries tend to have associated psychological issues. Simply going out in public may be emotionally difficult, and going for mental health help may be even more challenging. This problem is compounded by geographical barriers that affect even soldiers without physical injuries. Specifically, the veteran population is disproportionately located in rural areas and smaller towns with limited health care resources. According to a recent article in *Psychiatric News*, “Although only 20 percent of Americans live in rural areas, 41 percent of the patients getting their care through the VA live in sparsely populated regions.” Appropriate medical infrastructure for these veterans does not exist within a reasonable distance, and mental health services are especially hard to come by in rural areas [3]. These soldiers still need such care, as well as access to a range of therapies and support groups, but the perceived and real difficulties of getting that care may prove overwhelming.

1.1 Rationale

In our search for novel solutions to address some of the issues preventing soldiers from getting care, we looked at new forms of technology, and decided to focus on today’s popular social networking/game platforms called Virtual Worlds (VWs). Virtual world technologies have several affordances that might surmount obstacles to care; they are typically anonymous (as people log in with a fictitious name) and therefore help mitigate perceived stigma, do not pose a challenge to most people with disabilities (as long as they can use a computer), can connect people with others of common interest over broad distances (as they are networked), and are persistent and accessible to anyone with a computer and a broadband connection.

The Coming Home project chose the free-to-use virtual world *Second Life*™ (SL) as a primary platform of investigation because of its size, large number of users, and expanding scripting language that is useful for designing applications within the virtual world, as well as communicating with external applications. It may also have some inherent appeal for younger soldiers as a game-like platform. According to a recent study, younger veterans (18-24) also face more risk for mental health concerns than do more experienced soldiers [4].

There were two goals in mind when developing an application in this virtual world environment. The first was to create an engaging activity that could help alleviate stress. The second goal was to introduce concepts from proven systems of stress amelioration that are also being explored within the virtual world by ICT, such as Mindfulness Based Stress Reduction and Yoga. One concept that fit both criteria was controlled breathing. Potential physical health benefits due to controlled breathing have been indicated by research showing that regular breathing accompanied by biofeedback for as little as ten minutes a day can be effective in lowering blood pressure [5].

Research from Stanford’s Virtual Human Interaction Lab (VHIL) led us to hypothesize that modeling an exercise behavior could potentially have psychological benefits. VHIL’s research has shown that avatar usage manifests behavioral changes

via what they have termed the “Proteus Effect.” Participants in a VHIL study who observed their avatar exercising reported significantly higher levels of exercise in the physical world following their session than those who watched a representation of someone else exercising, or watched themselves not exercising [6]. Research by Lim and Reeves [7], also at Stanford, shows that the act of allowing the user to choose their avatar, and viewing that avatar in a 3rd person perspective, leads to greater arousal and an increased sense of presence in a virtual environment, which correlates with increased engagement. These studies are influential in our thinking that activities done in the virtual world with one’s avatar may be an important means for affecting behavioral changes in the physical world.

We combined these two fields of research, biofeedback with controlled breathing and exercise with the Proteus Effect, to produce a virtual jogging activity within *Second Life*TM. To go beyond the typical way people cause an avatar to run (keyboard and mouse clicks), we designed a virtual jogging path where control of the avatar is done via steady, rhythmic breathing. The avatar’s movement itself is a feedback mechanism indicating success at the activity: the avatar progresses from standing to walking to running depending on how long the user has successfully completed the activity. The user only influences the movement of the avatar; the direction of travel is handled automatically by scripts written in Linden Scripting Language (LSL) that guide the user’s avatar around a circular path circumnavigating a virtual island.

2 Design

Although rhythmic breathing was used as the control mechanism for the virtual jogging activity, we eschewed use of a spirometer for this implementation, although the SL viewer could be modified to be compatible with such a device. Using a spirometer would have required more costly alteration of the SL viewer, and would have also required additional maintenance to accommodate the frequent operational changes in the official SL platform. More importantly, it would have limited the target audience to those who had personal use of a spirometer. We wanted our implementation to be usable without any specialized equipment to reach the largest number of people. We chose to use an ordinary microphone that most *Second Life*TM residents already used for voice chatting to others in the world. The microphone was used to “hear” the sound of the breath rather than measure the amount of air exhaled.

This novel approach of using the sound of rhythmic breathing to move an avatar was made possible by seldom used speech functions provided by *Second Life*TM, which allow the use of a microphone to be detected, returning three levels of sound amplitude categorized as *low*, *medium*, and, *high* in value. Rather than use these functions to directly trigger animations or sounds, we altered them to transmit the information about which sound levels were detected via private chat channels accessible only by LSL scripts.

An LSL algorithm evaluates these *low*, *medium*, and *high* values and compares them to the time they were detected. By calibrating input settings for the microphone to return a *high* level when exhaling, the starting point of each exhalation can be determined. The length of time the participant sustains the exhalation can be determined by assuming that *high* values of low incidence indicate exhalation, while assuming that multiple *low* values or values of high incidence indicate interruption.

The speech functions do not continuously return values if the level of sound detected remains the same, so a comparison of recent values and their times detected is necessary.

This comparison process is also made crucial by the necessity of implementing error correction strategies: Initial testing with computer generated tones of constant amplitude revealed that the information received is not completely reliable as occasional glitches are encountered which briefly return lesser values before again returning the correct value. Values can become even more unreliable in practice because the sound of exhaling or the sound of air hitting the mic may not remain consistent. In addition to comparing incidence of values, strategies to exclude these deviations rely on proper calibration of the mic to return high values when exhaling (but not so high as to produce audio clipping), having a quiet room with little extraneous sound, and the use of headphones to block sounds from SL reaching the mic. The need for error correction is further compounded by the fact that there is no way to know if the volume level suddenly drops to zero, such as by accidentally or purposely muting the microphone. Because of this, we use “timeout” function that assumes nothing is happening if values have not changed after a sufficient amount of time.



Fig. 1. The green bar rises and falls in response to the length of time volume is detected by the user’s mic, providing the user real time feedback. Matching one’s breathing to the pace of the red bar causes the avatar to run. The green icons above the user’s head are SL’s default indication that the microphone is in use.

The process of making the avatar move involves the user attempting to match a slow and steady “guide breath” that is both audibly and visibly represented. (See Figure 1) The audible guide breath heard by the user mirrors the rate of rise and fall of a visible, red “guide bar” that appears on the user’s interface. The user’s own breath is visually indicated by a green bar that rises and falls in response to their detected exhalation, providing real time feedback. When the user exhales during appropriate exhalation period as indicated by the guide bar, their avatar will begin to move along the predefined course. Values returned outside this exhalation period are ignored.

2.1 Additional Challenges

Choosing the *Second Life*TM platform for its strengths also meant tackling its constraints. Regular users often choose to override the default animations because they seem awkward, and there are LSL animation override solutions for this provided by third party developers. We also chose to use a custom set of animations that more closely resemble jogging. In addition to customized animations, we wanted to provide visual feedback in the form of progressively changing running speeds. Normally, the avatar in SL can only move at two different speeds. No solution existed to change this constraint because it is hard-coded into the platform itself, whose code exists on remote servers inaccessible to the end user, and cannot be directly altered.

Our solution to this challenge was to first have the avatar sit on a primitive geometric object, also known as a “prim.” Using LSL, the prim is fully transparent, and made to navigate automatically through the world at a rate corresponding with animations matching various speeds of movement as triggered by the user’s breath input. This process gives the overall appearance that the user’s avatar walks and runs through the world at speeds not regulated by the *Second Life*TM platform.

With the final functionality in place, we were finally able to set up a study to determine the effectiveness of a breath-activated jogging scenario.

3 Study

We conducted a preliminary study with 27 participants that measured mood and arousal effects produced by engaging in this virtual jogging scenario. The participants were a mixed group of male and female subjects, though not specifically selected from the veteran population that is the intended beneficiary of our research. Veterans may already be suffering from psychological and health problems, and we therefore reasoned that a future study should be pursued with them only if random participants of a preliminary study demonstrated measurable mood and arousal effects that were desirable.

All participants for this study used a male avatar wearing Army fatigues, and were tested at the same virtual location within *Second Life*TM. All users were given verbal instructions and used the same pair of headphones and microphone for the study. The avatar was viewed in 3rd person perspective.

Three instruments administered prior to and after the virtual jogging activity were used to measure the effects on mood and arousal states. These were the arousal section of the *Pleasure, Arousal, and Dominance* (PAD) scale, and the *Positive Affect*

(PA) and *Negative Affect* (NA) scales. We found participants' experience using the jogging path had effects on these measures, resulting in a significant decline in mean score for all three scales. The significance of the change was: .004 for PAD, .015 for PA, and .006 for NA. Given the significant findings for the overall scales, post-hoc item analyses were conducted to determine the specific nature of the change participants experienced. When these were done for the ten items of the PA, there were significant (drops) findings for two of the items ($p < 0.05$), *interested* and *inspired*. One item that showed a more significant decline ($p < 0.01$) was *enthusiastic*. Post-hoc analyses on the ten items of the NA found significance ($p < 0.05$) for one item, *distressed*, and greater significance ($p < 0.01$) for two items, *nervous* and *upset*.

4 Discussion

Overall, this indicates that this virtual jogging activity, in its totality, tends to help participants feel more relaxed and calm. Cognitively, there is a lessening of arousal, trending toward feeling more sluggish and dull. While there is a decline in some positive emotions such as inspired and enthusiastic, there was a significant decline in the negative emotions as well, notably in the amount of distress, nervousness, and upset that the participants reported. This decline in negative emotions may have notable implications for those with disorders such as anxiety and stress. We believe that the results may have implications for new avenues of research in the field of Resilience Training, an area of extreme interest to the military [8]. We also expect our intended audience (veterans) will experience greater psychological effects via the Proteus Effect because future users will create and use their own avatars and will therefore feel an association with the avatar as a projection of themselves, while the preliminary users did not.

Acknowledgments. The project described herein has been sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM). Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

1. Milliken, C.S., Auchterlonie, J.L., Hoge, C.W.: Longitudinal Assessment of Mental Health Problems Among Active and Reserve Component Soldiers Returning From the Iraq War. *JAMA* 298(18), 2141–2148 (2007)
2. RAND Corporation (RAND Health Division and Rand National Security Research Division): Invisible Wounds of War: Psychological and Cognitive Injuries, Their Consequences, and Services to Assist Recovery (2008),
<http://rand.org/pubs/monographs/MG720/>
(retrieved June 12, 2008) from The Rand Organization Website:
<http://www.rand.org/multi/military/veterans>
3. Levin, A.: Vets in Rural Areas Face Multiple Barriers in Care. *Psychiatric News* 42(10), 12 (2007), American Psychiatric Association
<http://pn.psychiatryonline.org/content/42/10/12.full>

4. Karen, H., Seal, K.H., Bertenthal, D., Miner, C.R., Sen, S., Marmar, C.: Bringing the War Back Home: Mental Health Disorders Among 103 788 US Veterans Returning From Iraq and Afghanistan Seen at Department of Veterans Affairs Facilities. *Arch. Intern. Med.* 167(5), 476–482 (2007)
5. Grossman, E., Grossman, A., Schein, M.H., Zimlichman, R., Gavish, B.: Breathing-control lowers blood pressure. *Journal of Human Hypertension* 15(4), 263–269 (2001)
6. Fox, J., Bailenson, J.N.: Virtual self-modeling: The effects of vicarious reinforcement and identification on exercise behaviors. *Media Psychology* 12, 1–25 (2009)
7. Lim, S., Reeves, B.: Being in the Game: Effects of Avatar Choice and Point of View on Arousal Responses During Play. Paper presented at the International Communication Association, Dresden, Germany (2006)
8. Wadsworth, S.M., Riggs, D. (eds.): See, e.g. *Risk and Resilience in U.S. Military Families*. Springer, NY (2011)

Low Power Wireless EEG Headset for BCI Applications

Shrishail Patki¹, Bernard Grundlehner¹, Toru Nakada², and Julien Penders¹

¹ Holst Centre/imec, High Tech Campus 31, 5656 AE Eindhoven, The Netherlands

² Panasonic Corporation, Japan

{shrishail.patki, bernard.grundlehner, julien.penders}@imec-nl.nl,
nakada.toru@jp.panasonic.com

Abstract. Miniaturized, low power and low noise circuits and systems are instrumental in bringing EEG monitoring to the home environment. In this paper, we present a miniaturized, low noise and low-power EEG wireless platform integrated into a wearable headset. The wireless EEG headset achieves remote and wearable monitoring of up to 8 EEG channels. The headset can be used with dry or gel electrodes. The use of the headset as a brain computer interface is demonstrated and evaluated. In particular, the capability of the system in measuring P300 complexes is quantified. Applications of this prototype are foreseen in the clinical, lifestyle and entertainment domains.

Keywords: Brain computer interface, EEG, headset, low power, wireless, wearable, ASIC.

1 Introduction

Among the bio-potential signals, acquiring EEG signals is particularly challenging because of their small amplitude (typically between $1\mu\text{V}$ - $20\mu\text{V}$) and wide frequency range (typically from 0.1Hz to 100 Hz). The system needs to have a low noise floor to record such small amplitudes. Monitoring EEG in ambulatory environment is becoming more important not only in clinical domains but as an extra parameter for various life-style, brain computer interface (BCI) and entertainment applications. In order to address a wide variety of applications, it is important to have a system which is miniaturized, wearable, wireless and provides flexibility and comfort to the user.

The use of dry electrodes presents an attractive option for a wearable system due to their quick and easy setup, but is associated with several challenges [4]. Dry electrodes cause high electrode offset, have higher impedance than the gel electrodes and imply additional constraints on the amplifier design such as high CMRR to reject the power line interference amplified by the use of dry electrodes.

Several wireless EEG headsets have been introduced in the last few years. Commercially available headsets such as Emotiv EPOC[2] can work continuously for 12 hours while measuring a total of 14 channels including EMG and EOG with EEG. The information recorded with the headset is then used to develop gaming and life-style applications. Neurosky Mindset [9] uses one dry contact resistive electrode to record EEG from FP1 position to determine mental state of the user. Emband from EmSense [3] uses dry electrodes in a headset to measure emotional and cognitive

activity which is further used for market research. The B-Alert headset [1] from Advanced Brain Monitoring determines the different states of a user by measuring the EEG and combining it with algorithms for compensating motion artefacts. This headset provides 12 hours of autonomy with 2 Li-ion batteries of 500mAh.

These systems have had a significant impact in making EEG available to application research. The power consumption however remains too high to enable long-term use, especially with wireless transmission. Strategies to reduce system level power consumption while maintaining good signal quality are required. This paper builds upon our earlier work on wireless EEG monitoring [7], and presents the development of a miniaturized wireless sensor node focused on low noise and low power and its integration into a wearable headset with dry electrodes. The wireless headset is benchmarked against a reference wired system [5](g.USBamp from g.tec) on several aspects, starting from technical quantities (noise floor, frequency bandwidth, SNR) to more application-oriented quantities (similarity measures), finally comparing BCI performance with the oddball paradigm [8] based on P300 response.

2 Wireless EEG Headset

2.1 Low-Power Wireless EEG Electronics

The wireless EEG system consists of four blocks: a proprietary eight channel ASIC (Application Specific Integrated Circuit), a microcontroller, a radio and a power circuit. The eight channel ASIC [11] is an ultra low-power solution for small sized and autonomous EEG acquisition systems. The channels are sequentially digitized by an 11-bit analog digital converter (ADC). The ASIC also features impedance monitoring circuitry to measure electrode impedance up to 33 K Ω . During EEG acquisition, the ASIC consumes 240 μ W. The sensor node as shown in Fig. 1. (left panel) was designed by separating the analog and digital electronics, in order to reduce the interference from the radio communication on the ASIC. The analog module includes the ASIC, power management circuit and an impedance measurement circuit. The digital module comprises a microcontroller from Texas Instruments (TI MSP430F1611), a radio from Nordic Semiconductors (nRF24L01+) and an optional grounding switch capability for switching off unused channels. The microcontroller with its five low power modes is optimized to achieve extended battery lifetime in portable measurement applications. The microcontroller controls the digital settings of the ASIC, provides a 32.768Hz clock to the ASIC and handles the start-up sequence for the ASIC. The ASIC transfers the samples digitally to microcontroller through the SPI interface. Then the samples are processed and transferred to the radio through a second SPI interface.

In addition to the built-in impedance measurement functionality of the ASIC, an impedance circuit is included to have a larger dynamic range. This extends the impedance measurement range to the case of dry electrodes. Additionally, a mechanism that issues a trigger with every application related event is implemented in the sensor node to synchronize the acquired EEG data in case of a BCI application.

The sensor node measures 35mm x 25mm x 5mm excluding the 140mAh Li ion battery. It can measure up to 8 channels of EEG at a maximum sampling frequency of 1024Hz. The data is transmitted wirelessly to a computer for real time display, storage and analysis.

2.2 Power Optimization Strategy

Several power optimization techniques have been applied to reduce the power consumption on hardware and firmware levels. These can be summarized as follows:

- All the selected components have low quiescent and leakage current.
- Parts of the system (e.g. impedance measurement) are switched off when not required.
- The interrupts are used to wake up the microcontroller from low power modes; Direct Memory Access (DMA), which does not require CPU intervention, is used to transfer data from the ASIC to the microcontroller memory and to the radio. This strategy allows keeping the microcontroller in low power mode and thus reduces the power consumption.
- The radio is duty cycled.

2.3 Integrated Wireless EEG Headset

The wireless sensor system is integrated into a wearable headset for easy and fast monitoring of EEG, as shown in Fig. 1. (right panel).



Fig. 1. System electronics (left) wireless headset (right)

The electrodes are placed at the pre-defined positions of C3, Pz, C4 and Cz with the reference and bias electrodes behind the ears. The use of dry electrodes allows quick and easy user setup. The electrode position Pz is chosen specifically to address the needs of Brain Computer Interface (BCI) applications. The electrodes integrated in the headset are commercially available reusable Ag/AgCl coated contact resistive electrodes [10], with special contact posts and do not require the use of gel. Nevertheless, the headset design offers the possibility of gel injection for applications where low contact impedance is required. The electrode assembly with the metallic holder and spring provides many degrees of freedom for better adaptation to the user's head.

3 Evaluation Methods

3.1 System Characterization

The wireless sensor node is characterized from a system perspective by measuring input referred noise, signal to noise ratio and power consumption. The impedance measurement circuitry is then calibrated against reference resistances. Input referred noise is calculated by short circuiting all the channels to reference and bias electrode. For calculating signal to noise ratio, a pre-recorded clean EEG signal is played back using a DAC board from National Instruments [6]. This signal is then re-recorded with the wireless sensor node. The ratio between the original signal to the difference between original and re-recorded signal is calculated as the signal to noise ratio. The same test is performed with a sine wave of amplitude 200 μ Vpp at 10Hz.

3.2 Frequency Response Comparison

This test consists of comparing the frequency response of measured EEG data, over a frequency range of 1Hz-30Hz, recorded with the headset and the reference system simultaneously, both with eyes open and eyes closed. The wireless headset is used with dry electrodes, while gel (Ag/AgCl) electrodes and skin preparation are used with the reference system. The measurement electrode was placed at C3 position with the reference and bias electrodes at left and right mastoid respectively. Electrodes from both the systems were placed very close to each other avoiding any contact. EEG was recorded for the duration of 1 minute, each for eyes open and eyes closed. The data was sampled at 256Hz and 1024Hz with reference system and headset respectively. The data was analyzed in frequency domain using Welch's power spectral density estimation method, using a Hann window with 50% overlap, and a spectral resolution of 0.125 Hz. For all four subjects (male, Holst Centre), the spectral energy between 1 and 30Hz of the recordings of eyes open and eyes closed was compared using a Pearson product-moment correlation coefficient. This comparison is based on standardized data, for which the standard score (z-score) calculated by subtracting the mean (μ) from the signal and dividing it by the signal's standard deviation (σ).

3.3 P300 Response Extraction and BCI Performance

An application based on the oddball paradigm was used for comparing the wireless headset with the reference system in their capability of extracting P300 response. Simultaneous recording were performed with the measurement electrode at Pz and reference and bias electrodes at left and right mastoid respectively. In order to enable the synchronization between the two systems, the subject was asked to bite before every experiment. These tests measure the user's P300 response to visual stimuli with an oddball and item selection task.

Oddball Task with Visual Stimuli. Subject's response during the display of target (a circle shape, 'o') or control (a cross shape 'x'), was recorded in this experiment. The experiment protocol used was as follows:

- The target or control was made visible for 200ms with the probability of target set at one fourth. The time interval between each flash was set at 150ms.
- This procedure was repeated 20 times and considered as a single set.
- Four such sets with each of the four subjects were recorded.
- The user was asked to think 'Yes' immediately after the appearance of target.
- Corresponding to every visual stimulus, a trigger signal was issued to synchronize the EEG with application related screen flashes.

Item Selection Task. This experiment consists of visual stimuli in the form of flashes that are associated with list of four different items. The items were flashing in random order and the user's response to designated items was recorded. The protocol was as follows

- The user was informed about the designated item on a screen which was visible for 4 seconds.
- Then each item was flashed 5 times in a random fashion. The flashing frequency was set at 200ms and a time interval of 150ms was set between each flash.
- The user was requested to think 'Yes' immediately after appearance of designated item.
- There were 20 flashes in total. This forms a single set.
- Five of such sets were recorded.
- Corresponding to every visual stimulus, a trigger signal was issued to synchronize the EEG with application related screen flashes.

The data obtained from headset and the reference system is compared by calculating similarity between the two signals and selection accuracy for the item selection task. The recorded data was first band pass filtered using a 1Hz-10Hz digital band-pass filter. Additionally, the maximum of the covariance function in an interval [-1, +1] second around the tooth clicks was used to synchronize the EEG data recorded with the headset and the reference system, after which the application-specific triggers, as generated by the headset, can be used also for the EEG data recorded with the reference system. Epochs of EEG data were then extracted 100ms prior to each trigger with the total duration being 700ms. Each of these 700ms epochs of EEG were averaged with respect to each symbol or item. The data which includes large noise components such as eye movement or an eye blink was discarded and not used for analysis. The epochs of EEG data as recorded with the headset were compared with the data from the reference system by computing the Pearson product-moment correlation coefficient for each epoch.

The item selection task also includes the calculation of the selection accuracy. The accuracy was calculated by averaging 5 data sets of EEG for each trial and each item and discriminating between target and control. The discrimination was based on Linear Discrimination Analysis (LDA) classifier. For the computation of classification accuracy Leave-One-Out cross validation was applied on all data belonging to one subject.

4 Results and Discussion

4.1 System Characterization

The signal to noise ratio (SNR) of 25dB from Fig. 2 suggests that the system can detect very small signal components which are 25dB lower than the average RMS value. For a full scale sine wave of 10Hz, SNR is 40dB. The input referred noise (Fig. 2) shows a similar frequency response than the EEG ASIC, which means the frequency characteristics remain consistent between the wireless sensor node and the ASIC.

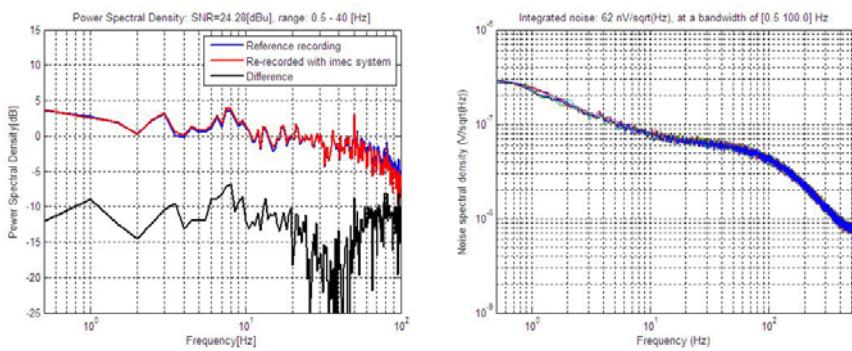


Fig. 2. Signal to Noise ratio for pre-recorded signal and input referred noise

The impedance (Fig. 3) calibration plot indicates that impedance is within the $\pm 10\%$ range over complete dynamic range. The power consumption in Table 1 denotes that power can be optimized efficiently and provides more flexibility to user for varying parameters in case of different applications. Input referred noise is measured as well, to track the possible degradation in noise performance as the radio transmission frequency changes.

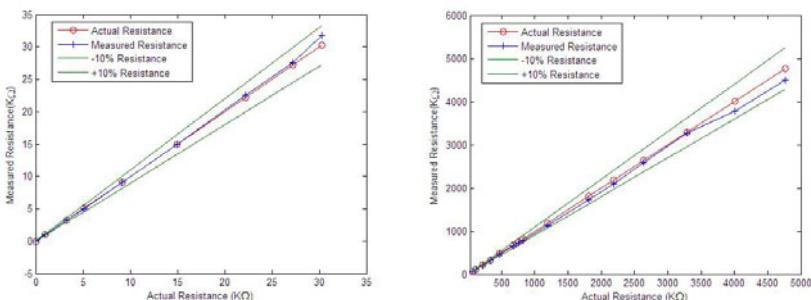


Fig. 3. Impedance measurement over complete dynamic range from $1\text{K}\Omega$ to $4.7\text{M}\Omega$

Table 1. Power Consumption Optimization

Sampling Freq (Hz)	Transmission Freq (Hz)	No. of Channels	Impedance	Current (mA)	Power (mW)	Noise (nV/ $\sqrt{\text{Hz}}$)	Autonomy on 140mh battery (hours)
1024	1024	8	ON	4.65	13.95	62	30
1024	1024	8	OFF	3.56	10.68	62	39
1024	512	8	OFF	3.07	9.21	62	45
512	512	8	OFF	2.27	6.81	62	61
512	256	8	OFF	2.08	6.24	62	67
256	256	8	OFF	1.68	5.04	68	83
256	128	8	OFF	1.53	4.59	68	91
256	36.6	1	OFF	1.11	3.33	70	126

4.2 Frequency Response Comparison

The frequency distribution as shown in Fig. 4 indicates high correlation between the two systems. This fact is reflected in Table 2 displaying correlation coefficients between two systems.

Table 2. Correlation coefficients for eyes open/ eyes closed

Subject	Eyes Open	Eyes Closed
Subject 1	0.98	0.84
Subject 2	0.94	0.94
Subject 3	0.98	0.97
Subject 4	0.95	0.96

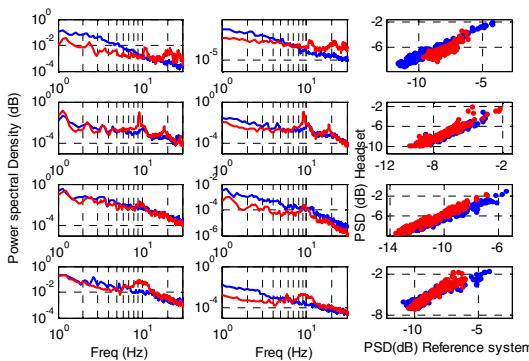


Fig. 4. PSDs of EEG, recorded with Eyes open(blue) and Eyes closed (red) with reference system (column 1) and headset (column 2), for each subject (rows). The third column shows scatter diagrams of the normalized PSD data (z-scores) of the headset against the PSD data of the reference system, recorded with eyes open (blue) and eyes closed (red).

In addition it also indicates successful monitoring of alpha wave brain activity (8-12Hz) for all the subjects.

4.3 P300 Response Extraction and BCI Performance

A histogram of all correlation coefficients ($n = 4700$) for the EEG data around every visual stimulus is displayed in Fig. 5. The median of all coefficients is 0.75, thus suggesting that a significant linear relation exists between the headset and the reference system.

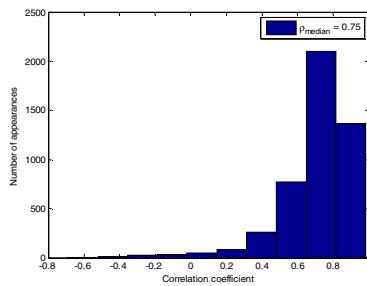


Fig. 5. Correlation Coefficient between headset and reference system

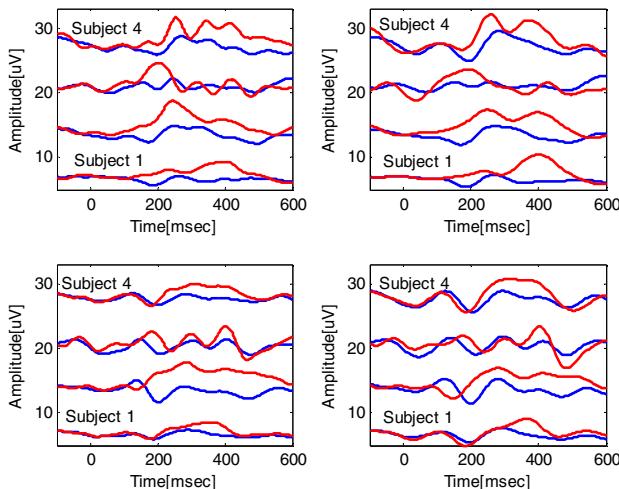


Fig. 6. P300 response for all the 4 subjects during oddball task (top row) and item selection task (bottom row) from reference system (right column) and headset (left column). The blue denotes control while red denotes target.

It can be observed from Fig. 6 that the degree of correlation is very high between the two systems for detecting P300 response. At the same time, the P300 component is relatively weak in both systems. This leads to quite low accuracy of the item

selection task (BCI) except for subject 2, as can be seen in Table 3. This could be due to the significant human factor involved which make the result depending on the person's ability to make a selection strongly. However, it can also be observed that the scores of the two systems are very similar, which suggests that the headset is equally capable of a control task as the reference system.

Table 3. Item Selection Accuracy Comparison

Subject	Wireless Headset	Reference System
Subject 1	70%	62.5%
Subject 2	93.75%	100%
Subject 3	77.5%	70%
Subject 4	53.1%	50%

5 Conclusion and Future Work

As a step towards ambulatory EEG monitoring, the development of a miniaturized, low noise and low power system and its integration into a wearable headset is presented. The wireless sensor node consumes 9.21mW when measuring 8 channels of EEG at 1024Hz, providing autonomy of 45 hours with a rechargeable battery of 140mAh. Power consumption drops to 3.3mW for a single-EEG channel recording at 256Hz sampling frequency, leading to over 5 days autonomy on the same battery. The SNR is 25dB when measured on real EEG signals, and 40dB for a sine wave of amplitude 200 μ Vpp at 10Hz. The signals measured with wireless headset show high correlation with signals recorded using commercially available wired reference system, with respect to frequency response of EEG waves and P300 response extraction. The headset uses off-the-shelf, dry electrodes, which drastically improve ease-of-use and set-up time.

Future work shall look at embedding algorithms for artefact filtering and feature extraction. Artefacts could range from physiological signals like EOG, EMG or motion and electrode artefacts. Feature extraction may include frequency, spatial and temporal analysis. Trade-offs between local extraction and wireless data streaming must be investigated. Furthermore, the design of dry electrodes and active electrode front ends will be explored to improve signal to noise ratio. Breakthroughs in ultra low power circuits will further decrease the power consumption enabling smaller systems with higher autonomy.

Acknowledgement. The authors would like to thank Jef Van de Molengraft, Refet Firat Yazicioglu and the Body Area Network team at imec / Holst Centre for their contribution to this paper.

References

1. B-Alert, <http://www.b-alert.com>
2. Emotiv, <http://www.emotiv.com>
3. EmSense, <http://www.emsense.com>

4. Ruffini, G., Dunnea, S., Fuentemillab, L., Graub, C., Farrés, E., Marco-Pallarés, J., Watts, P.C.P., Silva, S.R.P.: First human trials of a dry electrophysiology sensor using a carbon nanotube array interface. *Sensors and Actuators* 144, 275–279 (2008)
5. Gtec, <http://www.gtec.at>
6. National Instruments. Ni pxi4461, <http://sine.ni.com/nips/cds/view/p/lang/en/nid/13634>
7. Brown, L., Van de Molengraft, J., Penders, J., Yazicioglu, R., Torfs, T., Van Hoof, C.: A low-power, wireless, 8-channel EEG monitoring headset. In: IEEE Engineering in Medicine and Biology Society Conference (2010)
8. Squires, N., Squires, K., Hillyard, S.: Two varieties of long-latency positive waves evoked by unpredictable auditory stimuli in man. *Electroencephalogr. Clin. Neurophysiol.* 38(4), 387–401 (1975)
9. Neurosky, <http://www.neurosky.com>
10. Integra Neurosupplies. E021, <http://integralife.com/>
11. Yazicioglu, R.F., Merken, P., Puers, R., Hoof, C.V.: A 200uw eight-channel acquisitionasic for ambulatory eeg systems. In: IEEE Int. Solid-State Circuits Conf., pp. 164–165 (2008)

Virtual Mouse: A Low Cost Proximity-Based Gestural Pointing Device

Sheng Kai Tang, Wen Chieh Tseng, Wei Wen Luo, Kuo Chung Chiu,
Sheng Ta Lin, and Yen Ping Liu

User Experience Design Department, ASUS Design Center, ASUSTeK Computer Inc.,
15, Li-Te Rd., Peitou, Taipei 112, Taiwan
{tony1_tang, parks_tzeng, hunter_luo, patrick_chiu,
ted1_lin, ping1_liu}@asus.com

Abstract. Effectively addressing the portability of a computer mouse has motivated researchers to generate diverse solutions. Eliminating the constraints of mouse form factor by adopting vision-based techniques has been recognized as an effective approach. However, current solutions cost significant computing power and require additional learning, thus making them inapplicable in industry. This work presents the Virtual Mouse, a low-cost proximity-based pointing device, consisting of 10 IR transceivers, a multiplexer, a microcontroller and pattern recognition rules. With this embedded device on the side of a laptop computer, a user can drive the cursor and activate related mouse events intuitively. Preliminary testing results prove the feasibility, and issues are also reported for future improvements.

Keywords: Proximity based device, IR sensor, pointing device, virtual mouse.

1 Introduction

According to statistics of the International Data Corporation (IDC), the global consumption of laptop computers has passed the “Gold Cross” for the first time, exceeding that of desktop computer in 2010. This phenomenon reflects the importance of “portability” among computer users [2][3].

However, in addition to a touch pad or a track point already embedded in a laptop computer supporting pointing tasks, carrying an additional pointing device, i.e. a full size computer mouse, to enhance performance and ergonomics is inevitable and inconvenient. This additional device is owing to that the event structure of a touch pad or a track point requires two fingers, mostly a thumb and a forefinger, acting awkwardly to activate a drag action or a drag-selection action; meanwhile, that of a computer mouse needs only one finger to activate such action. This feature lowers the efficiency and comfort of a touch pad or a track point over those of a conventional computer mouse.

Effectively addressing the portability of a computer mouse has motivated industrial designers to flatten a computer mouse for easy carry [2] and even to slot into the laptop body while not in use [3]. Conversely, computer scientists create a computer

mouse without a physical body to achieve ubiquitous computing. These invisible mice can translate hand gestures and movements into mouse events by using computer cameras as a signal input [1][5][6][7].

Still, those inflatable mice can not fulfill stringent the ergonomic requirement of intensive operations. In contrast, despite eliminating concern over ergonomic constraints, vision-based approaches expend a significant amount of computing power, i.e. equal to almost a high performance GPU, to recognize predefined hand gestures, which increases the mental load of user, thus making these solutions inapplicable in industry.

This work, describes a novel proximity-based pointing device consisting of 10 pairs of inexpensive infrared transceivers, a multiplexer, a microcontroller and a pattern recognition algorithm. This embedded device on the side of a laptop computer detects the intuitive hand-movements of users on the tabletop and further translates them into mouse movements and events (Fig. 1). Equipped with full mouse functions without a physical mouse body, the proposed device is referred to as a Virtual Mouse, similar to terminology used in previous works.

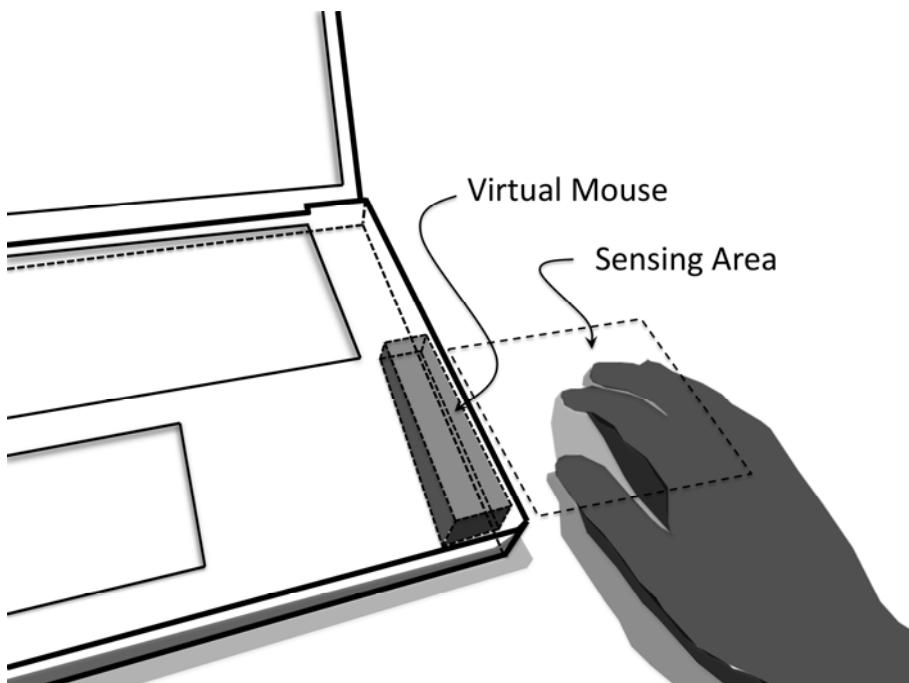


Fig. 1. Concept of Virtual Mouse

2 Implementation

2.1 Cost Savings Infrared Transceiver

The arrangement of transmitter and receiver for an infrared (IR) transceiver, i.e. a well-developed device in the market, allows it to detect an object and determine the

distances to it accurately. IR transceiver is thus extensively adopted as a reliable input device in security, robotics and home automation.

Rather than purchasing these mature products ranging from US\$ 30 to 100 or even more expensive ones designed for specific purposes, researchers without an electronic engineering background can still easily assemble components to construct an IR transceiver in order to resolve diverse laboratory problems and explore new sensing possibilities. Therefore, this work presents a simple IR transceiver rapidly by using only an IR LED, phototransistor, capacitor and two resistors, which cumulatively cost less than US\$ 0.5.

Specifically, a 3mm IR LED is powered through a 330-Ohm resistor, while a 3mm phototransistor is powered through a 20K-Ohm resistor with 5-Volt DC power supply. Additionally, the Base pin of the phototransistor is connected to an additional 0.1u capacitor for stabilization. The Base pin allows us to acquire linear signals within 6cm range, which is sufficient for our Virtual Mouse prototype (Fig. 2).

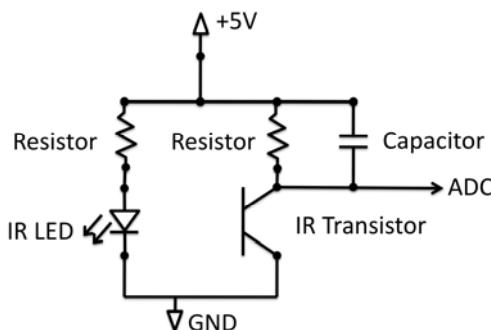


Fig. 2. Circuit scheme of homemade IR transceiver

2.2 Infrared Transceiver Bar

The customized IR transceiver provides a one-dimensional sensing ability. Combining 10 identical IR transceivers and arranging them in parallel allow us to create a device capable of detecting objects on a 4cm * 6cm two-dimensional plane. The shape of an object or its movement can be recognized after analyzing the sensor signals. Restated, this customized device is nearly equal to a touch pad that enables finger touch and gesture recognition.

Ten IR transceivers require a prohibitively expensive 10 analog input-pins of a microcontroller to read signals, explaining why the proposed device uses a multiplexer (MUX) as a digital switch to reduce the number of input pins. Therefore, 10 transceivers are connected to the MUX and the MUX is connected to a microcontroller as a de-multiplexer (Fig. 3).

2.3 Pattern Recognition

Based on 10 IR transceiver signals, this work also develops sequential rules to recognize diverse signal patterns, which are fundamental to driving a mouse cursor

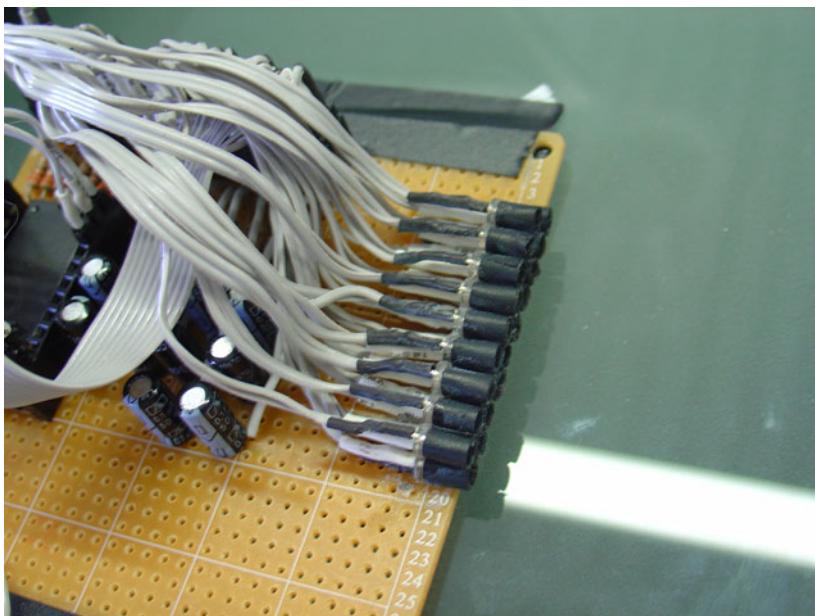


Fig. 3. Prototype of IR sensor bar

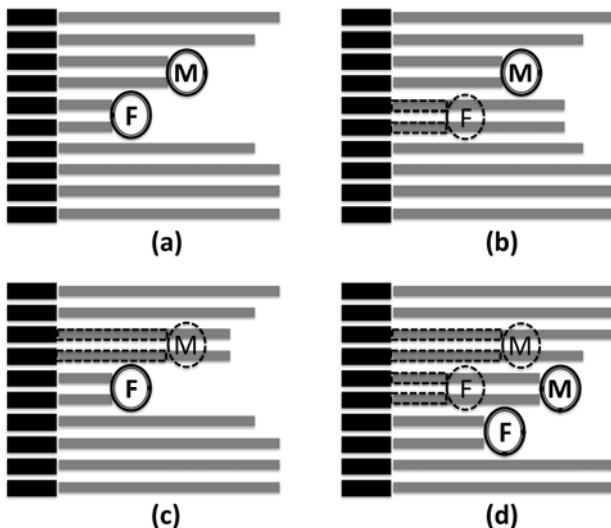


Fig. 4. Pattern recognition rules

and triggering corresponding events by hand, i.e. button down, button up, click, double click. In contrast with training models to achieve a high performance directly, this set of rules originates from observation of invited users and is intended mainly for rapid proof of concept to facilitate the future development involved in additional resources, e.g., software and firmware engineers.

Ten subjects, i.e. 5 male and 5 female, were invited to collect hand gesture patterns. Subjects were instructed to perform 6 actions within the sensing area, i.e. vertical move, horizontal move, diagonal move, forefinger click, forefinger double click and middle-finger click. Sensor signals were further recorded and analyzed.

Eventually, 4 rules derived from the previous 6 testing actions are placement, forefinger-up, middle-finger-up and move. A pattern in which the signal pattern is divided into two stages and the stage values subsequently decrease is recognized as placement, implying that the middle finger and forefinger appear in the sensing area (Fig. 4-a). A pattern in which the value of the second stage increases and exceeds that of the first stage is recognized as forefinger up (Fig. 4-b). If the value of first stage increases and exceeds that of the placement pattern, this pattern is recognized as middle finger up (Fig. 4-c). A pattern in which the value of second stage changes horizontally or moves vertically (or both) in comparison with that of the previous pattern is interpreted as move (Fig. 4-d).

2.4 Finite State Machine

Based on the above rules, a finite state machine (FSM) is designed to interpret hand gestures and trigger their corresponding mouse events. Consider a drag action, in which a touch pad or a track point requires two fingers to activate. The proposed FSM begins with the none-detection state (N). While the placement pattern is recognized, the FSM moves to ready state (R). While the forefinger-up pattern and placement are subsequently detected within 100 milliseconds, the FSM moves to left-button-down state (LBD) and triggers the left-button-down event. Notably, the FSM goes to ready

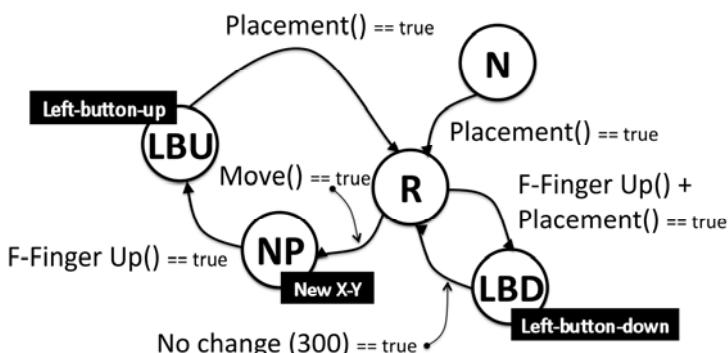


Fig. 5. Finite state machine for drag action

state (R) again if no new pattern is detected within 300 milliseconds. At this moment, while the move pattern is recognized, the FSM moves to new position state (NP) and triggers new X-Y coordinate event. While forefinger-up is detected, the FSM moves to left-button-up state (LBU) and triggers the left-button-up event. Once the placement is recognized, the FSM returns to ready state (R). Rather than using a computer mouse, the above sequence completes a drag action with hand and our Virtual Mouse (Fig. 5).

3 Preliminary Testing

Ten subjects invited for a previous observation used the virtual mouse again to complete tasks in a 480px * 360px simulation window. The simulation window has three 40px * 40px squares: one is blue, another is red and the other is transparent with a dashed outline. Subjects were requested to click and double click the red square and, then, drag the blue square to the transparent square sequentially 10 times. Subject performances were recorded for later analysis.

Analytical results indicate that the average completion rate of a click (84%) is higher than that of a double click (70%) (Fig. 6). Specifically, incomplete tasks initially occur several times, a phenomenon attributed to the required learning period. A double click has a 300 milliseconds time constraint making it more difficult to get used to than with a single click and also requiring a longer time to learn.

Our results further demonstrate that the total average completion rate of a click (77%), single plus double, surpasses that of a drag (70%) (Fig. 6). Given the lack of a specific trend on the charts, subjects were interviewed to identify potential reasons. Most subjects indicated that the sensing area of the Virtual Mouse prototype is insufficiently large. Their finger were thus always out of boundary when dragging; in addition, the FSM lost signals before the controlled square could reach the target area.

	Female					Male					
	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	AVG
Click	8	7	8	9	8	9	8	9	9	9	84%
Double Click	6	6	7	7	7	8	7	6	8	8	70%
Drag	8	8	9	7	7	6	7	7	6	7	72%

Fig. 6. Statistic result for preliminary user testing

4 Potential Applications

With this IR transceiver bar, two of them can be embedded at two sides of a laptop computer. For ordinary mouse functions, a right hand/left hand (according to user's

handedness) can easily drive the cursor and trigger events (Fig. 7-a). While using two hands simultaneously for manipulation, a user can achieve scale and rotation, which resemble those of a multi-touch pad and display (Fig. 7-b).

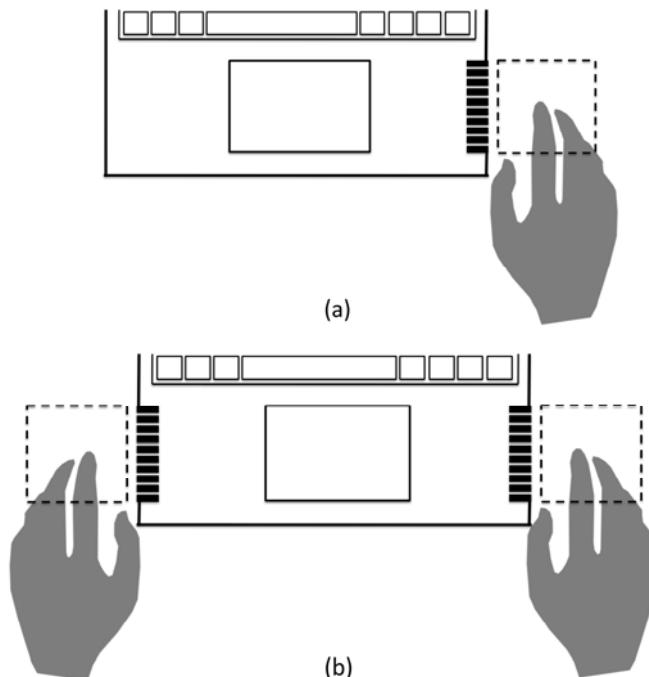


Fig. 7-a, 7-b. Ideas of potential application

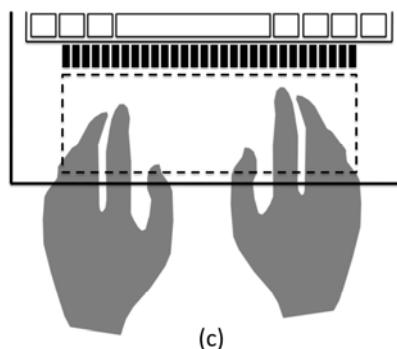


Fig. 7-c. Ideas of potential application

A longer IR bar can also be embedded at the upper edge of a palm rest, which sends IR signals to the lower edge of the palm rest. With such an arrangement, this IR bar encompasses the entire area of the palm rest and turns it into a sensible surface.

A user can perform all actions described above freely on the palm rest. Importantly, no additional plane area outside the laptop is required for operation, thus making the Virtual Mouse applicable under all circumstances (Fig. 7-c).

5 Related Works

To address the portability of computer mouse, researchers with industrial design background has developed volume-adjustable mice, e.g., Jelly Click [2] and Inflatable Mouse [3]. Jelly Click is a piece of soft plastic bag with a circuit board attached. Users are required to blow it up for use and flatten it for carry. Inflatable Mouse consists of a balloon-like inflatable structure. It can be a flat shape or a ready-to-grasp shape depending on the volume of machine-injected air. Users still need to carry the Jelly Click although it is flattened as thin as a piece of paper. Conversely, the Inflatable Mouse can be stored flat in a card slot of a laptop computer.

Computer scientists have adopted vision-based approaches to totally eliminate the constraints of mouse form factor and create invisible pointing devices, e.g., Visual Panel [7], Visual Touchpad [6], Hands Free Mouse [4] and virtual mice [1][5]. Visual Panel employed an arbitrary quadrangle-shaped panel and a tip pointer to realize a point device, whereas Visual Touchpad detected a fixed plane and finger tips to enable multi-touch. Hands Free Mouse simulated mouse clicks by simple hummed voice command, while head movements tracked by a webcam drove the cursor. Robertson et al.'s virtual mouse was a kiosk recognizing predefined hand signs to track hand movements. Conversely, Gai et al.'s virtual mouse recognized and tracked features of a scene captured by a camera of a mobile phone, thus turning camera motion into virtual mouse control.

6 Conclusions and Future Study

The Virtual Mouse has received considerable attention in both academic and industrial communities. Rather than focusing on a novel concept, the proposed Virtual Mouse prototype provides a cost-savings approach for mass production. Replacing the common vision-based scheme with an inexpensive IR sensor bar has cumulatively reduced the cost from \$US 50 to \$US 5.

Instead of adopting predefined gestures to activate mouse functions, emulating the intuitive finger gestures of conventional mouse to eliminate the learning curve is another benefit of this work. Unlike a camera collecting complicated imagery data for whole-hand gestures, our homemade IR sensor bar allows us to acquire adequate signals for subtle finger gestures.

Although issues such as increasing the smoothness and enlarging the sensing area require further improvement, this work significantly contributes to efforts to demonstrate the feasibility of a potential solution and develop technical specifications.

In addition to replacing the 3mm LED with a SMD one to increase the resolution and enlarge the sensing area, efforts are underway in our laboratory to modify the pattern recognition rules and link the signal to the operation system. Additional developmental results and evaluations will be published in the near future.

References

1. Gai, Y., Wang, H., Wang, K.: A virtual mouse system for mobile device. In: Proceedings of the 4th International Conference on Mobile and Ubiquitous Multimedia, pp. 127–131. ACM Press, New York (2005)
2. Jelly Click, <http://www.designodoubt.com/entry/Jelly-click--mouse-for-laptop>
3. Kim, S., Kim, H., Lee, B., Nam, T., Lee, W.: Inflatable mouse: volume-adjustable mouse with air-pressure-sensitive input and haptic feedback. In: Proceedings of CHI 2008, pp. 211–224. ACM Press, New York (2008)
4. Polacek, O., Mikovec, Z.: Hands free mouse: comparative study on mouse clicks controlled by humming. In: Proceedings of CHI 2010, pp. 3769–3774. ACM Press, New York (2010)
5. Robertson, P., Laddaga, R., Kleek, M.V.: Virtual mouse vision based interface. In: Proceedings of the 9th International Conference on Intelligent User Interfaces, pp. 177–183. ACM Press, New York (2004)
6. Shahzad, M., Laszlo, J.: Visual touchpad: a two-handed gestural input device. In: Proceedings of the 6th International Conference on Multimodal Interfaces, pp. 289–296. ACM Press, New York (2004)
7. Zhang, Z., Wu, Y., Shan, Y., Shafer, S.: Visual panel: virtual mouse, keyboard and 3D controller with an ordinary piece of paper. In: Proceedings of the 2001 Perceptive User Interfaces, pp. 1–8. ACM Press, New York (2001)

Innovative User Interfaces for Wearable Computers in Real Augmented Environment

Yun Zhou, Bertrand David, and René Chalon

LIESP laboratory, Ecole Centrale de Lyon, 36, avenue Guy de Collongue,
69134 ECULLY cedex, France
{Yun.Zhou,Bertrand.David,Rene.Chalon}@ec-lyon.fr

Abstract. To be able to move freely in an environment, the user needs a wearable configuration that is composed of a set of interaction devices, which allows the interaction at least one hand free. Taking into account the location (physical, geographical or logical) and the aimed activities of the user, the interaction style and devices must be in appropriate relation with the context. In this paper, we present our design approach and a series of real proposals of wearable user interfaces. Our research is investigating innovative environment dependent and environment independent interfaces. We describe these interfaces, their configurations, real examples of use and the evaluation of selected techniques.

Keywords: One-hand interaction, wearable interface, augmented reality, pico projector, context awareness, finger tracking.

1 Introduction

Ubiquitous computing and pervasive systems are important evolutions of information technology (IT) allowing new use in everyday life. The term ubiquitous computing was introduced by Mark Weiser [1], which focuses on the integration of technologies into daily life with the aim of binding the user, environment and technologies as one. The objective of ubiquitous computing is to eliminate the use restriction constraining the users to access the IT system only with fixed or portable computers and their classical graphical interfaces (GUIs), with WIMP style and devices (e.g., screen, keyboard and mouse). On the contrary, the wearable computer allows acting in mobility and in the context related to real environment [2]. The real environment can be augmented consciously to support the relation between real and virtual (digital) worlds. The works of Wellner [3] and Milgram [4] are pioneer works in the field of augmented reality environment. We have been evolving in this research field in relation with mobility for several years which can be characterized by two acronyms: MOCOCO (MObility, COntextualization and COoperation) and IMERA [5] (French acronym for Mobile Interaction with Real Augmented Environment). The real environment augmentation can be more or less conscious and can be conscious passively or actively. Recognition by the IT system of objects, actors or situations of interest without markers is the case of passive and unconscious augmentation. The other case is the augmentation with the use of passive or active markers. In the first

case, the IT system can discover these markers and use them in the process of treatment. In the second case, the active markers (e.g., the active RFID “Radio Frequency Identification” stickers) can address the IT system according to their own decisions. The IT system can for its part either be deployed in the environment with its sensors, or be dependent on the user interaction devices which build a unique relationship between real environment and IT system. In this paper, we are mainly concerned with this last approach: conscious augmentation by using passive markers. In our previous studies, we elaborated a referential [6] of interaction devices and a method to assist the choice of wearable computer interaction styles and devices in relation with the application field tasks and actions, in order to determine an appropriate set of devices for each actor. In this way, we mainly studied the use of hand free highly mobile and mixed reality mobile actor interactions [7].

In this paper, we present a series of innovative wearable interfaces based on the webcam capture, with the aim of allowing the user to have at least one hand free in the augmented reality environment in the context. We are concerned mainly with light, wearable and cheap user interfaces.

2 Related Work

In recent years, there have been a great number and variety of marker based interactions [8, 9, 10] that have made it possible to use contextual markers in a mobile environment. The CyberCode system [9] is a tagging system designed for the augmented reality, which is based on CyberCode. The PerZoovasive [11] learning environment is an adaptive and context-aware system which provides the support for learners in mobile pervasive situations by using QR codes (Quick Response codes). In addition, compared with other detection technologies such as RFID, the tag of AR-Toolkit (or QR code) is more economical. Our approach is inspired by these contextual markers, which can bridge the digital and real world in a light and cheap way. In order to interact easily with dynamic information we also attempted to integrate the use of a wearable projected interface based on a pico projector. Kurata et al. present the BOWL ProCam [12] that proposes the interaction techniques effectively employing both nearby projection surfaces such as the user’s hands and the far projection surfaces such as a tabletop and a wall. But they pay close attention to the technique on how to understand where the nearby-and-far-away surfaces are located, rather than the experience of the mobile projection of the users. The technology of Skinput [13] enables the skin to be used as an input surface, and their approach provides an always available, naturally portable, and on-body finger input system. However the projected interface is limited to a small size, which cannot enable more information to be presented. Our work is concerned with projected interfaces which can present information on a larger surface to view and interact with. The SixSense project [14] proposes a wearable device by superimposing the projected information onto the surfaces in the real environment. It focuses on gestures including the gestures supported by multi-touch systems, the freehand gestures and iconic gestures. However, it is a necessity for an interface to provide the menus and buttons to select in relation with the acquisition of the information. Our mobile wearable AR (Augmented Reality) system focuses on providing the user with instantaneous contextual information.

3 Light Mobile User Interfaces

We have designed and implemented a series of prototypes of innovative interfaces, which allow the user to interact with his/her environment with at least one hand free. These different prototypes are mainly based on: a webcam for perceiving the context and user interactions, and a goggle-attached small screen for visualizing text, image and video; otherwise, for larger field of vision and interface, the goggle attached screen is replaced by a pico projector, for dynamic information projection in the context. A small computing device is located in the user's pocket or his/her backpack. The software for these prototypes was developed on the Microsoft Windows platform by using C, C++, OpenCV, GTK and AR-Toolkit.

3.1 An In-environment, Fix Interaction Support

We are distinguishing mobility and nomadism. For us mobility is with a wearable computer and nomadism without, i.e. in the second case the environment is providing interaction support. In this way, we fixed the webcam on a plastic bracket on the desk of the workplace (Figure 1, a, b), or the wall of a bus shelter. With the help of this configuration based on the webcam capture of interaction and large screen or wall video projection we can establish an in-environment interaction. We studied the feasibility and effectiveness of the wearable camera interaction technique with a finger selection technique on a passive grid. The principle of the interaction is based on x-y coordinate detection of the finger position. We segmented the whole visual field of webcam into rectangular zones and related each zone with a unique event. The user can trigger asked action through selecting the related zone. The program processes the real-time video stream data which is captured by the webcam and tracks the positions of the coloured sticker located on the index finger. In a test application called "an intuitive dictionary" the user can directly search for the meaning of the words or the corresponding pictures. The paper grid support has been divided into 12 areas. Six words and six pictures are printed on the grid. When the user puts his/her finger over the corresponding areas of the words, a related picture appears on the screen. Similarly, the application also lets the user get the word by tapping on the picture.

Besides, we increased the number of zones. More and more grids have been put into the interface until the size of the zone is equal to the size of the green sticker. It was observed that the size of the zone could be reduced to be as small as the size of a finger sticker. This quick experiment indicated that the buttons and the markers, which we will mention later, can be recognized in a small size without disrupting the interaction. In order to enable the user to acquire the instant information contextually, we chose the markers which can be recognized by AR-Toolkit to link the real world and the digital one. The markers can be pasted on wall, books, newspapers, appliances or the doorplate (Figure 1, c, d). In this experiment, a fixed configuration was still used but with added markers integrated into each grid, and then the AR-Toolkit was used to recognize these markers. We mainly focused on the efficiency of marker recognition and their arrangement on the grid. To demonstrate the feasibility we implemented an application allowing the user to choose the information of interest from a list of book titles. Each book title has a corresponding marker printed aside it.

As the user selects the marker, he/she obtains the information about this book. We also employed the same method to study the passive marker size. It is proved that in a similar way the marker can be as small as the finger sticker.

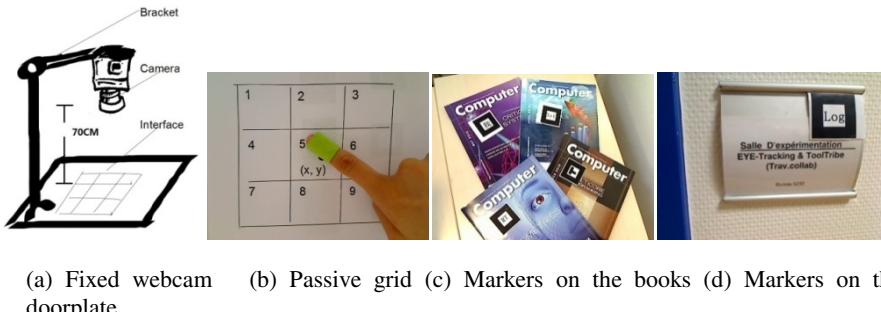


Fig. 1. Interaction in a fixed environment

3.2 Mobile Environment Dependent Interaction

In order to move from nomadic to mobile interaction (interaction based on a wearable computer) we abandoned the plastic bracket and dressed the user with a head-band attached webcam and on his/her goggles we attached a small screen. A laptop in the backpack allows the user to become free in a mobile situation (Figure 2, a, b). The selection feedback and task related information are presented on the small screen. The user can use the menus and markers located in the working or casual environment, mainly posted up on the wall in vertical position and at appropriate distance to make capture by the webcam on the head easier.



Fig. 2. Interaction in a mobile environment

Besides the finger selection, we also tested a mask based selection in order to present only one marker at a time to the webcam. The mask is the object which can shade other markers which can be in the form of a piece of paper, a hand-made stuff or even a notebook which can be flipped through (Figure 3, b, c). In the mask selection technique, the mask can be casually moved by the user. With an artificial mask, made like a switch, the user can flip up/down over the mask in the same way he/she can open or close a door. In this way a large set of markers located on a sheet can be used without recognition problems. Thus, we can obtain the dynamicity

allowing to take into account a large set of tasks. In this approach, the user can use in-environment fixed menus. In this case the advantage is to be in relation with the environment (i.e. contextual interaction) and the disadvantage is that these menu sheets can be spoiled by vandalism. Apart from two selection techniques as we stated above, we also introduced the page selection technique (Figure 3, d). We arranged only one marker on each page. The user can show the webcam one marker at a time by flipping through the pages. We also used a pack of cards in Rolodex® or flip card mode. Only one card can be shown at a time.



(a) Finger selection (b) Mask selection (c) Mask selection (d) Page selection

Fig. 3. Selection techniques

We wanted to improve the interface by enlarging the interactive surface and to provide a mobile environment independent interface. Our investigation applied a similar methodology as with the Goggle attached screen based prototype to study the projected mobile interface. In this case we replaced this screen by a pico projector which can project the interactive menus, the text, the image, the videos and other contents on any flat surface in the environment. In this way, the user, not only is freed from the limited screen size, but is also given an interaction space (Figure 2, c). We fixed the webcam on the user's body as an input device, and the projector as an output device. The webcam and the projector were combined together as a whole configuration, the position of which seriously affected the efficiency of the interaction. The fixing point of the configuration should be settled to capture the person's field of view and especially to distinctly and completely recognize the marker, in order to sustain the vibrations caused by the physical movement, to facilitate wearing, to observe the hand gestures and so on [12]. We evaluated the position of the fixing point in an empirical way by using "quick and dirty" observation, with the aim of finding an appropriate point on the basis of the small existing equipments; if the equipments had been wireless, as small and light enough as a button, perhaps we would have had more choices for locating it. Four points have been studied (Figure 2, d): the top of the head, ear side, the shoulder, and on the front of the chest. Wearing the configuration fixed on these four points respectively, the task of the user was to interact with the markers pasted on the wall. We tested the configuration by five evaluation components: the scale of the field of view of the webcam, the view stability, the view flexibility, the distortion of the webcam and projector caused by the angle and the facility of the fixation on the body. After observing the user's actions, we found that it was not easy to fix well or in a stable way on the shoulder or on the front of the chest, particularly when person is walking, mainly because the longest side of the pico projector used was orthogonal to the body. With regard to the position in front of the chest, if the markers are plastered equal to or higher than the line of sight, the webcam must be tilted

upwards to support the capture, which leads to distortion. Furthermore, if the height of the user is below the average height, he/she only can adjust the angle of the configuration to match the visual surface. Besides, the experiment showed that the height of the visual surface depends on the height of the eyes. Finally, we decided, in relation with previous discussion, to fix the webcam and the projector on the top of the head, by using a bike (or old ice hockey) helmet.

4 Scenarios and Applications

We created several scenarios and so implemented the relevant applications in order to present how these prototypes could promote interaction in the context and facilitate the mobile user to access to the information freely in a specific environment or a general setting. Two types of interfaces are proposed on the basis of the relation with the environment: an environment dependent interface and an environment independent interface. On the one hand, we define the environment dependent interface as the interface which is in strong relation with the in-environment information and markers (on walls, on doors, appliances or any other surfaces). This interface has the ability to provide intuitive interaction techniques which can recognize and understand the situation of the user and the real environment around him/her. The supports of information, that are the tangible markers, are static and protected against vandalism. In this way, the public or professional guiding information can be used for contextualization based on webcam recognition. On the other hand, we also designed an environment independent interface with which the user can acquire the right information at any time in any place. The contextualization is done by the user himself/herself through the way of showing appropriate contextualizing indications to the webcam. These indications can be grouped on a menu grid and selected by finger, by a mask or by flipping the pages of a notebook. On each flip card or Rolodex® card, we can find a textual and selected description of action to be done. In this way, the user can contextualize his/her working environment. With the pico projector based independent interface, the user can project the menus, schedules, websites, videos and other information on a plane surface (such as the wall) in the environment or on a small personal projection surface (such as a sheet of paper, a cardboard or a part of the human body). The latter case is mainly a way to solve the problem as usually the user cannot find an appropriate surface to project in a public place, as mentioned in [15]. The user is completely free to move his/her working environment and then can obtain contextual information independently from the environment. The scenarios and applications were created, to compare selection techniques and appreciate their usability.

4.1 Environment Dependent Interfaces

The first scenario is an indoor way-finding instruction application which can assist the user to find the destination, when he/she comes into a new building for the first time, and wants to find the office of a laboratory. Usually, the logos of the lab are fixed next to the entrance gate of the building. The markers are pasted on one side of the logo and can be read by the webcam configuration, and then identified to extract the contextual instruction for the indoor way-finding option. When he/she points to the

right side of the marker with his/her finger with a coloured sticker, an interface with the video information or the image instruction will pop up on the small screen attached on the goggles or be projected on the wall. Moreover, when the user reads a newspaper, he/she can only read the text and the image. However, through the pico projector and the markers, he/she can watch a vivid video augmented on the newspaper (Figure 4, a). In another more sophisticated interactive application called a “Research Team Interaction System”, the user can interact with a piece of paper containing markers and grids. When the user arrives at the lab, wants to ask the teacher a question and he/she door, he/she can consult the teacher’s schedule to visit him/her again. He/she will then look for the appropriate time by using the interface pasted on the door. First, he/she selects the action “to consult the timetable”. Next, he/she clicks the photo of the teacher and chooses the date of interest. Finally, the user views the timetable and decides the next time he/she can come again (Figure 4, b). In some public places, like at the bus stop or the gas station, the user can get useful advice about nearby restaurants appropriate for their needs. The location, opening hours, a telephone number can be proposed to the users through the markers. In addition, when the tourist is waiting for the bus, if he/she suddenly wants to know the local weather forecast for that day, he/she only needs to select the marker and read the information. In this way, the user can obtain the instant information quicker and more easily than by getting the information with several text inputs and menu selections through his/her phone.



(a) Video augmentation (b) Dependent interface (c) Independent interface (d) Selecting the person of interest

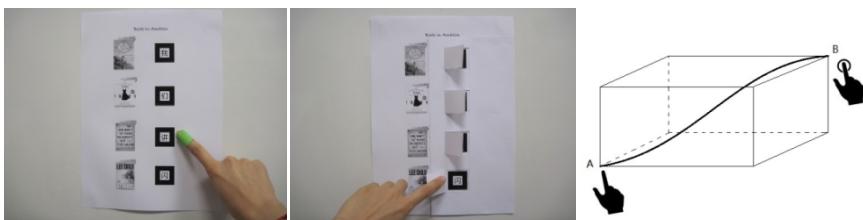
Fig. 4. The user selects the content of interest to project

4.2 Environment Independent Interfaces

Apart from the general environment, the applications are created according to the specific working situation, with an environment independent interface. In the case of maintenance activity in augmented reality we provide the user with the appropriate information in relation with his/her activities by using a pico projector. For example, in an industrial scenario, a novice technician needs to replace a system board of a laptop computer. The task steps are then projected and superimposed on the real objects. Then the technician contextually reads and views the sequence of actions and guidelines. This is not only the process of task completion, but also the process of mobile learning based on the context. Moreover, the pico projector allows more dynamic behaviours by projecting the information on the wall and then allowing the user to interact with it freely (Figure 4, c, d).

5 The Evaluation of Selection Techniques

We conducted an experiment to compare the finger selection technique with the mask selection technique on a small scale. The goal was to compare the efficiency of two selection techniques. The equipment of the experiment includes a book list which has four markers to select, a mask, and a green sticker which can be located on the index finger (Figure 5, a, b). The experimental task consists in selecting items of interest from a list of book titles. Each book title has a corresponding marker printed aside it. As long as the user selects the marker, he/she will obtain the information on the book, which will be shown on the goggle screen.

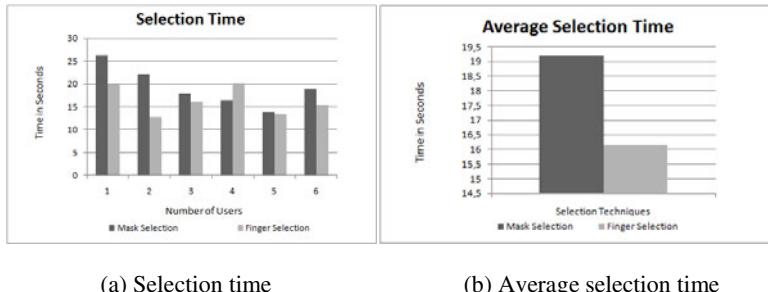


(a) Booklist selected by finger (b) Booklist selected by mask (c) The pathway of the finger

Fig. 5. Two selection techniques and the pathway of the finger in a space

We compared these two techniques by evaluating the selection time. Six volunteer participants performed five trials with each technique. Obviously, these two techniques take almost the same distance to reach the target (Figure 5, c), in spatial situation the distance is the length of curve between A and B. We defined that the point A belongs to the plane A and the point B belongs to the plane B. However, since the actions of opening and closing the mask cost extra time compared to the finger selection technique, we assumed that the selection time of the former is more than the latter. We added a timing counter into the programs to record the time. The selection time is composed of three parts: the reaction time, the internal decision time, and the execution time. The reaction time is the time between the appearance of the book list and the start of the action i.e. when the user starts to search an item in the book list. The internal decision time includes the process of searching for an item of interest and deciding the target item. The execution starts from the first movement of the finger or the mask, and ends when the user clicks (finger selection technique) or presents (mask selection technique) the target marker. However, it is unreasonable to test the selection time alone. Thus we gave the subjects a task asking them to check the information beneath the four markers in a random sequence. We recorded the elapsed time of the task, instead of the selection time. The results (Figures 6, a, b) suggest that the selection time by finger is less than the selection time by mask except. The average selection time by finger and by mask is respectively 16.17 seconds and 19.2 seconds. Though the selection time by mask is more than with the finger, the results indicated that the users preferred the mask selection technique because they find it easier to flip the mask, compared with pointing at a target carefully with the index finger in a space. It has been also observed that two participants usually started from a point on plane A and

then moved to the target on plane B as the first item selection process, and then kept the finger on the plane B. When selecting the next item, they started from plane B each time until the whole process ended. However, the rest of the participants often started from a point in plane A, which entails a longer selection time.



(a) Selection time

(b) Average selection time

Fig. 6. Selection time and average selection time for mask and finger selection techniques

In order to improve our interaction techniques, we first plan to design the vision based hand gesture interfaces in our future work. Based on the results mentioned above, the principle is that the user can interact on the plane A instead of B. When the user wants to interact with the projected surface, he/she does not have to click the menus and icons of the surface; instead, he/she operates the interface in a more natural and intuitive way. Our aim is to eliminate the inconvenience of pointing at a target carefully. We will mainly focus on the gestures which require less learning, recalling and adaptation of the user, based on the tasks and context. In brief, our vision-based hand gesture techniques and interfaces will not only give the user a more comfortable interactive experience, but also allow the user to interact contextually.

6 Conclusion

This paper described our approach to explore innovative user interfaces and wearable configurations to enable the user to access to the information freely and with simple quick interaction techniques. We presented the design and implementation of our prototypes which are based on a camera capture and computer vision techniques. The wearable configurations are mainly composed of a webcam, a small screen attached to the goggles and a handheld computing unit. For more advanced, larger interaction output, the user can replace the screen by a pico projector. To test and demonstrate the usability of the mobile interactions in the context, we created several scenarios and implemented several applications, based on our different prototypes. Meanwhile, we also explored three selection techniques: finger selection, mask selection and page selection techniques. Finally, we evaluated the efficiency of mask and finger base selections. The results of the experiment lead us to propose another interaction technique combining finger selection and hand gestures. Though limits exist, our method and prototypes still have advantages. Our mobile AR system allows the interaction in mobility with, if possible, at least one hand free. With the marker, the

user can obtain the information in the context, in a simple and quick way. Besides, more dynamic interactions can be achieved through the projector, as with complementary contextual information which can be achieved by active or passive RFID stickers to obtain “Proxemic Interactions” [16].

References

1. Weiser, M.: The Computer for the Twenty-First Century. *Scientific American*, 94–101 (1991)
2. Plouznikoff, N., Robert, J.M.: Caractéristiques, enjeux et défis de l'informatique portée. In: Proceedings of IHM 2004, pp. 125–132 (2004)
3. Wellner, P., Marckay, W., Gold, R.: Computer Augmented Environments: Back to the Real World. Special Issue of Communications of the ACM 36(7) (1993)
4. Milgram, P., Drascic, D., Grodski, J.J., Restogi, A., Zhai, S., Zhou, C.: Merging Real and Virtual Worlds. In: Proceedings of IMAGINA 1995, Monte Carlo (1995)
5. David, B.T., Chalon, R.: IMERA: Experimentation Platform for Computer Augmented Environment for Mobile Actors. In: 3rd IEEE International Conference on Wireless and Mobile Computing, Networking and Communications, WiMob, pp. 51–57. IEEE Computer Science, New York (2007)
6. David, B.T., Champalle, O., Masserey, G., Chalon, R.: From Task Model to Wearable Computer Configuration. In: Winckler, M., Johnson, H. (eds.) TAMODIA 2007. LNCS, vol. 4849, pp. 261–266. Springer, Heidelberg (2007)
7. David, B.T., Masserey, G., Champalle, O., Chalon, R., Delotte, O.: A Wearable Computer based maintenance, diagnosis and repairing activities in Computer Augmented Environment. In: Proceedings of EAM 2006, Valenciennes (2006)
8. Hornecker, E., Psik, T.: Using aRToolKit markers to build tangible prototypes and simulate other technologies. In: Costabile, M.F., Paternó, F. (eds.) INTERACT 2005. LNCS, vol. 3585, pp. 30–42. Springer, Heidelberg (2005)
9. Rekimoto, J., Ayatsuka, Y.: CyberCode: Designing Augmented Reality Environments with Visual Tags. In: Proceedings of DARE 2000 on Designing augmented reality environments. ACM Press, Elsinore, Denmark (2000)
10. Rouillard, J.: Contextual QR Code. In: Proceedings of the 2008: the 3rd International Multi-Conference on Computing in the Global Information (ICCGI 2008). IEEE, Los Alamitos (2008)
11. Rouillard, J., Laroussi, M.: PerZoovasive: contextual pervasive QR codes as tool to provide an adaptive learning support. In: Proceedings of 5th international conference on Soft computing as transdisciplinary science and technology. ACM Press, France (2005)
12. Kurata, T., Sakata, N., Kourogi, M., Okuma, T., Ohta, Y.: Interaction Using Nearby-and-Far Projection Surfaces with a Body-Worn ProCam System. In: Proceedings of The Engineering Reality of Virtual Reality 2008 in the 20th Annual IS&T/SPIE Symposium on Electronic Imaging, pp. 6804–6816 (2008)
13. Harrison, C., Tan, D., Morris, D.: Skinput: Appropriating the Body as an Input Surface. In: CHI 2010, Atlanta, USA (2010)
14. Mistry, P., Maes, P., Chang, L.: WUW-Wear Ur World-A Wearable Gestural Interface. In: CHI 2009, Boston, USA (2009)
15. Ko, J.C., Chan, L.W., Hung, Y.P.: Public issues on Projected User Interface. In: CHI 2010 (2010)
16. Greenberg, S., Marquart, N., Ballendat, T., Diaz-Marino, R., Wang, M.: Proxemic Interactions: the New Ubicomp? *Interactions XVIII*(1), 42–50 (2011)

Part V

Avatars and Embodied Interaction

Influence of Prior Knowledge and Embodiment on Human-Agent Interaction

Yugo Hayashi¹, Victor V. Kryssanov¹, Kazuhisa Miwa², and Hitoshi Ogawa¹

¹ Faculty of Information Science and Engineering, Ritsumeikan University

1-1 Nojihigashi, Kusatsu 525-8577, Japan

yhayashi@fc.ritsumei.ac.jp, {kvvictor,ogawa}@is.ritsumei.ac.jp

² Graduate School of Information Science, Nagoya University

Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan

miwa@is.nagoya-u.ac.jp

Abstract. An experiment was conducted to capture characteristics of Human-Agent Interactions in a collaborative environment. The goal was to explore the following two issues: (1) Whether the user's emotional state is more stimulated when the user has a human schema, as opposed to a computer agent schema, and (2) Whether the user's emotional state is more stimulated when the user interacts with a human-like ECA (Embodied Conversational Agent), as opposed to a non human-like ECA or when there is no ECA. Results obtained in the experiment suggest that: (a) participants with a human schema produce higher ratings, compared to those with a computer agent schema, on the emotional (interpersonal stress and affiliation emotion) scale of communication; (b) A human-like interface is associated with higher ratings, compared to the cases of a robot-like interface and a no ECA interface, on the emotional (e.g., interpersonal stress and affiliation emotion) scale of communication.

Keywords: Embodied Conversational Agent, Human-Computer Interaction, User Interface.

1 Introduction

Recently, there is a popular trend in the IT industry to develop and deploy Embodied Conversational Agents (ECAs) that would facilitate collaboration of various system users by adopting new technologies for the user-system interaction. One of the important issues in this area is to understand cognitive and emotional characteristics of the ECA-mediated communication process (Nass & Steuer, Tauber, 1994). The underlying question is: what are the factors that influence these characteristics?

2 Related Work

In the initial stage of communication with a stranger, people usually rely on some prior knowledge about the conversation partner. Social psychology research has indicated the importance of the top-down processing, based on prior knowledge about

the speaker, such as 'schema' and 'stereotypes' in interpersonal cognition. It is usually assumed that people use reference to a schema to understand utterances during conversation with a computer agent (Fisk & Taylor, 1991).

Hayashi & Miwa (2009) conducted a psychological experiment, in which schemas were controlled to explore characteristics of communication, when human and computer agents coexist. Results obtained showed that the schema about the communication partner affects emotional characteristics of communication: participants experienced more positive emotions towards the partner when they believed that the partner is a human but not a computer agent. In Yamamoto et al. (1994), participants of an experiment played Shiritori, a popular Japanese word-game, with a partner using a computer. Even though the actual identity of the partner was a computer agent, the participants, who were misinformed that they were facing a human player, gave significantly higher pleasure ratings than those who were informed that they were facing a computer agent. The study also suggested that a schema about the partner affects emotional characteristics of communication.

How can we stimulate human schemas and thus influence emotional characteristics while communicating with ECA? Sproull et al. (1996) found that people respond differently to a talking-face interface compared to a text-only interface. It was pointed out that people are more likely to ascribe personality attributes to displayed faces, and have higher arousal when view faces while interacting with a talking-face interface, compared to a text-only interface. It remains, however, unclear whether the participants experienced the same kind of emotional states as during conversations with human partners.

3 Purpose of the Study

In the presented study, the examined hypothesis is that the more an ECA has human-like characteristics, the more the users would rely on a human schema and deem the agent as human. It is, therefore, expected that when users interact with a human-like ECA, they would have more positive emotions towards the partner, compared to the case of a non human-like ECA, or when there is no ECA. Specifically, the following two points will be investigated:

1. Whether the user's emotional characteristics are more stimulated when the user has a human schema, as opposed to an agent schema.
2. Whether the user's emotional characteristics are more stimulated when the user interacts with a human-like ECA, as opposed to a non human-like ECA, or when there is no ECA.

4 Methods

Experiments were conducted where two participants are engaged in a rule discovery task and communicate with each other using an Internet chat. The participants are required to find the sequence for the number of objects that appear on their computer displays. In Figure 1, there are a total of six objects presented in the center: four black and two white. Below the objects, there is a text field where the participants input and

receive messages. Just one sentence per trial is permitted, and at most 30 words are accepted. Buttons for changing objects, sending messages, and for terminating the experiment are placed at the bottom of the screen.

A conversational computer agent (see Figure 2) used in this study has an ability to meaningfully respond to sentences input by the participants. The agent first extracts keywords from the chat messages and then activates scripts for generating sentences, based on the graphical images displayed (for details on the agent design, see Hayashi & Miwa, 2009).

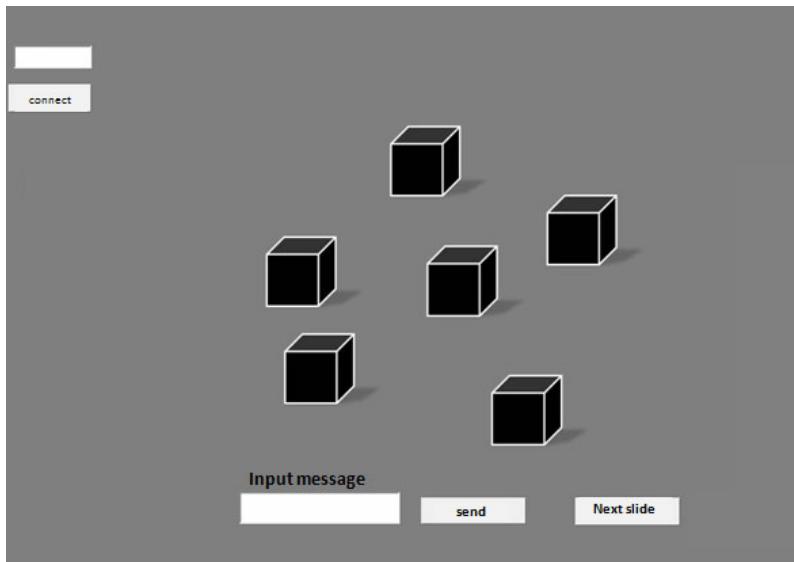


Fig. 1. Screen shot of the experimental setup

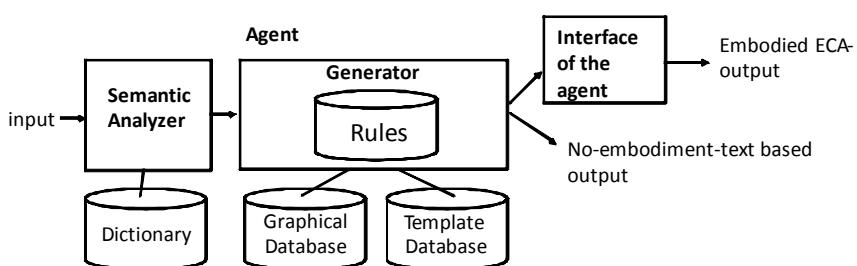


Fig. 2. Agent design

5 Experimental Design

In Experiment 1 (see Figure 3), psychological characteristics of participants having a human schema were investigated. The participants' behavior was controlled by telling

them if a human or a computer agent was supposedly involved into the collaboration. In reality, only a computer agent was used in the experiment. No interface agent (i.e. no embodiment) was deployed, and there was no ECA in both conditions. Participants having the human schema were labeled as “HUMAN condition”, and participants having the agent schema were labeled as “AGENT condition”.

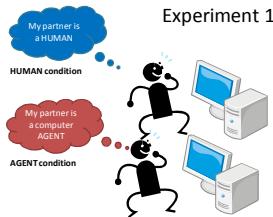


Fig. 3. Conditions of Experiment 1

Next, to investigate if a human-like ECA affects the same characteristics as found in Experiment 1, two conditions were added in Experiment 2. Results obtained were compared with those of the experimental participants under the AGENT condition. Two conditions in Experiment 2 were the same as the AGENT condition, excepting for the fact that the agents had a graphical representation and were visualized. In the human-like condition, there was a human avatar displayed, and in the robot-like condition, there was a robot avatar displayed (see Figure 4).

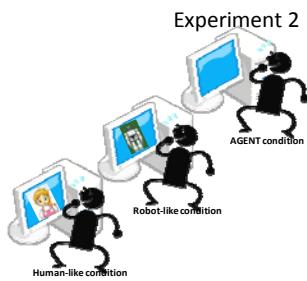


Fig. 4. Conditions of Experiment 2

49 university students (37 male, 12 female, average age=22.45 y.o.) voluntarily participated in the experiment, and each participant was randomly assigned to every condition.

6 Collected Data

The participants answered a questionnaire two times, before and after completing the experimental procedure (see Figure 5). The time allocated for the procedure was 5

minutes for each participant. The questionnaire used was originally proposed by Tsuduki & Kimura (2000), and it is designed to assess psychological characteristics of communication.

The questionnaire is comprised of 16 questions, each on a five-point scale. The 16 questions are classified into three groups to evaluate: (1) interpersonal stress, (2) affiliation emotion, and (3) information propagation. The first measure represents the “interpersonal stress” factor consisting of five characteristics about feelings, such as tension, severity, and fatigue. The second measure represents the “affiliation emotion” factor consisting of eight characteristics about feelings, such as friendliness, ability to discuss personal matters, and happiness. The third measure represents the “information propagation” factor consisting of three characteristics about feelings, such as purpose and effectiveness in collecting information. The first two groups are defined as “emotional characteristics”, and the third as “cognitive characteristics” of communication.

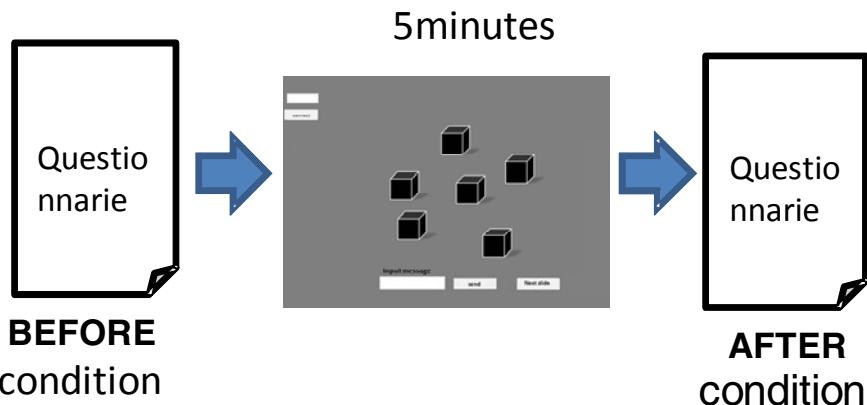


Fig. 5. Experimental procedure

7 Analysis

7.1 Experiment 1

Figure 6 shows results of Experiment 1. The vertical axis represents the mean value of the ratings obtained for the measures listed on the horizontal axis. A 2×2 ANOVA was conducted on each questionnaire measure with the instruction condition (HUMAN condition vs. AGENT condition) and the time condition (BEFORE condition vs. AFTER condition) as mixed-subject factors. There was no correlation detected between the two factors for all questionnaire measures, excepting for the questionnaire measure ‘speed of communication’ ($F(1,26)=5.54$, $p=.015$).

The analysis suggests that the influence of instruction were not changed over time for most of the questionnaire measures. Effects of instruction for each questionnaire measure were also investigated. Results obtained suggest that the conditions were

significant factors for measures, such as 'ease loneliness', 'happiness', 'feeling tension', 'feeling close', and 'expression of kindness' ($F(1,26)= 4.16$, $p=.052$; $F(1,26)= 3.97$, $p=.057$; $F(1,26)= 4.49$, $p=.043$; $F(1,26)=3.35$, $p=.079$; and $F(1,26)= 6.52$, $p=.017$, respectively). There were no significant differences for measures such as 'speed of

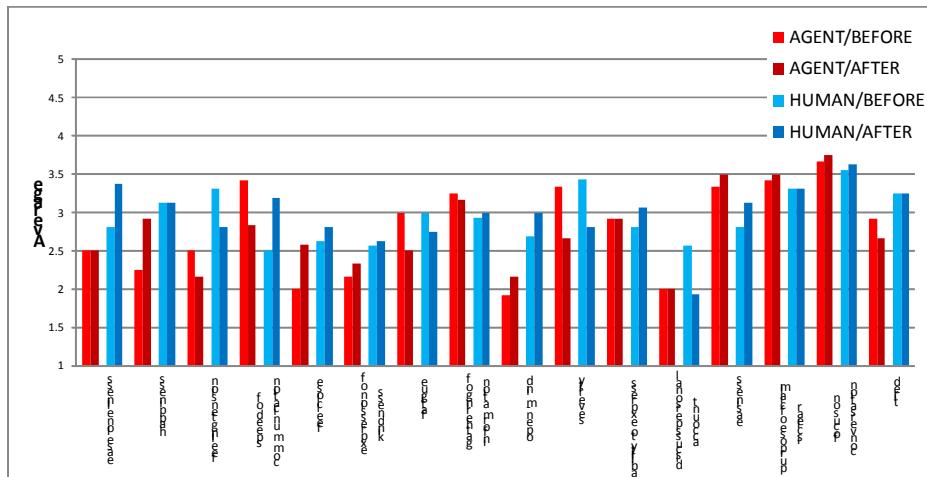


Fig. 6. Results of Experiment 1

communication', 'expression of kindness', 'fatigue', 'gathering of information', 'severity', 'ability to express', 'discuss personal account', 'easiness', 'purpose of claim is clear', 'focus on conversation', 'tired' ($F(1,26)= 1.18$, $p=.287$; $F(1,26)= 1.47$, $p=.237$; $F(1,26)= 0.11$, $p=.746$; $F(1,26)= 0.59$, $p=.451$; $F(1,26)= 0.1$, $p=.758$; $F(1,26)= 0.01$, $p=.932$; $F(1,26)= 0.74$, $p=.397$; $F(1,26)= 2.22$, $p=.149$; $F(1,26)= 0.24$, $p=.632$; $F(1,26)= 0.14$, $p=.716$; $F(1,26)= 1.95$, $p=.175$).

7.2 Experiment 2

Figure 7 shows results of Experiment 2 (the axes are labeled as in Experiment 1). A 3 x 2 ANOVA was conducted on each measure with the avatar condition (Human-like condition vs. Robot-like condition vs. AGENT condition with the same data as was used in Experiment 1) and the time condition (BEFORE condition vs. AFTER condition) as mixed-subject factors. There was no correlation detected between the two factors for all questionnaire measures, excepting for the measure 'severity' ($F(2,30)= 3.64$, $p=.039$).

To investigate the influence of embodied agents, main effects of the interface for each questionnaire measure were further analyzed. Results obtained indicate that the conditions were significant factors for measures, such as 'ease loneliness', 'happiness', 'feel close', 'expression of kindness', and 'severity' ($F(2,30)= 3.71$, $p=.036$; $F(2,30)= 6.62$, $p=.004$; $F(2,30)= 3.84$, $p=.033$; $F(2,30)= 3.72$, $p=.036$; and $F(2,30)= 3.3$, $p=.051$, respectively). Although, there were no significant differences for measures such as 'feeling tension', 'speed of communication', 'expression of kindness',

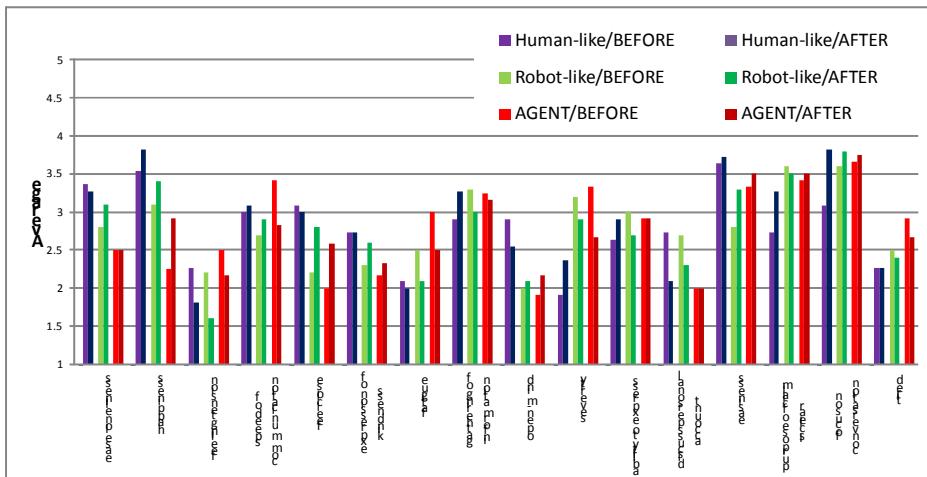


Fig. 7. Results of Experiment 2

‘fatigue’, ‘gathering of information’, ‘ability to express’, ‘discuss personal account’, ‘easiness’, ‘purpose of claim is clear’, ‘focus on conversation’, ‘tired’ ($F(2,30)=0.47, p=.632$; $F(2,30)=0.73, p=.49$; $F(2,30)=1.29, p=.291$; $F(2,30)=1.85, p=.175$; $F(2,30)=0.11, p=.9$; $F(2,30)=0.12, p=.9$; $F(2,30)=0.88, p=.425$; $F(2,30)=2.16, p=.133$; $F(2,30)=1.63, p=.212$; $F(2,30)=0.52, p=.598$; $F(2,30)=1.88, p=.17$).

A further analysis was conducted by using the Ryan’s method. Results obtained indicate that the conditions between the Human-like condition and the AGENT were significant factors for measures, such as ‘ease loneliness’, ‘happiness’, ‘feel close’, ‘expression of kindness’, and ‘severity’ ($p=.009$; $p=.001$; $p=.010$; $p=.020$; $p=.034$). On the other hand, results obtained indicate that the conditions between the Robot-like condition and the AGENT were only significant factors for measures, such as ‘happiness’ ($p=.037$).

8 Summary of Results

Table 1 summarizes results of the analysis of the major effects of instruction. In the table, the asterisk, plus, and minus signs stand for significant, marginal, and no statistical differences detected, respectively.

The results indicate that the detected effects of instruction are more significant for the “Affiliation emotion” and “Interpersonal stress” factors. This suggests that participants with a HUMAN schema produce higher ratings, compared to those with an AGENT schema on the emotional characteristics of communication scale.

Table 2 summarizes results of the analysis of the influence of the interface, where the asterisk, plus, and minus signs stand for significant, marginal, and no statistical differences detected, respectively.

Table 1. Summary of Experiment 1 findings

		HUMAN condition vs AGENT condition
Interpersonal stress	feeling tension	*
	severity	-
	fatigue	-
	easiness	-
	tired	-
Affiliation emotion	ease bneiness	+
	happiness	+
	feelcbose	*
	open-mind	-
	expression of kindness	*
	discuss personal account	-
Information propagation	focus on conversation	-
	speed of communication	-
	gathering of information	-
	ability to express animus	-
	purpose of claim is clear	-

Table 2. Summary of Experiment 2 findings

		Human-like condition vs AGENT condition	Robot-like condition vs AGENT condition
Interpersonal stress	feeling tension	-	-
	severity	+	-
	fatigue	-	-
	easiness	-	-
	tired	-	-
Affiliation emotion	ease bneiness	*	-
	happiness	*	*
	feelcbose	*	-
	open-mind	-	-
	expression of kindness	*	-
	discuss personal account	-	-
Information propagation	focus on conversation	-	-
	speed of communication	-	-
	gathering of information	-	-
	ability to express animus	-	-
	purpose of claim is clear	-	-

These findings indicate that the Human-like interface was rated higher, compared to the Robot-like and the no embodiment interface on the emotional scale of communication.

9 Discussion

9.1 Influence of Schema

In Hayashi & Miwa (2009), the authors used the same questionnaires and analyzed the effects of schema during participants' communication with a computer or human agent. In their study, when the participants were instructed that the partner is human, they produced more positive ratings towards the "Affiliation emotion" and

“Interpersonal stress” characteristics of communication. In the presented study, it was confirmed that participants with a human schema give more positive evaluations of the emotional characteristics of communication, compared to the participants with an agent schema.

In the previous study, the participants’ psychological characteristics were measured after the task. In the presented study, the participants’ psychological characteristics were assessed two times, and it was found that the effect of schema does not change over time. This suggests that the emotional characteristics emerge in a situation where participants have a consistent schema throughout communication with the partner.

9.2 Influence of Embodied Agents

The Human-like interface obtained higher ratings, compared to the Robot-like and to the no embodiment interface on the emotional scale of communication. This may indicate that the more ECA has human-like characteristics, the more the users would rely on a human schema and deem the agent as human. We found no previous studies that would focus on the effects of schema potentially affecting emotional characteristics of communication throughout an on-line conversation with an ECA.

There were studies on ECAs that focused on the creation of a realistic embodied agent [4]. The given work presents results that may be useful for the development of human-like conversation scenarios with a computer agent. The obtained results stress the importance of designing an ECA that would stimulate the user’s schema.

10 Conclusions

This study’s goal was to investigate the following two issues: (1) Whether the user’s emotional characteristics are more stimulated when the user has a human schema, as opposed to a computer agent schema, and (2) Whether the user’s emotional characteristics are more stimulated when the user interacts with a human-like ECA, as opposed to interacting with a non human-like ECA, or when there is no ECA. The study’s main findings can be formulated as follows: (a) participants with a human schema produce higher ratings, compared to those with an agent schema, on the emotional (interpersonal stress and affiliation emotion) scale of communication. (b) The human-like interface obtains higher ratings, compared to the robot-like and to the no embodiment interfaces, AGENT on the emotional (interpersonal stress and affiliation emotion) scale of communication. In social psychology, it is usually considered that schema and emotion co-occur in social interaction. Our results support this hypothesis, and also provide a new insight: when one fosters a human schema with the interface, this may enhance the user’s positive emotional states.

References

1. Fisk, T.S., Taylor, E.S.: Social cognition. McGraw-Hill Education, New York (1991)
2. Hayashi, Y., Miwa, K.: Cognitive and emotional characteristics in Human-Human/Human-Agent Interaction. Journal of Human Interface Society 10, 445–456 (2008)

3. Hayashi, Y., Miwa, K.: Cognitive and Emotional Characteristics of Communication in Human-Human/Human-Agent Interaction. In: Jacko, J.A. (ed.) HCI International 2009. LNCS, vol. 5612, pp. 526–531. Springer, Heidelberg (2009)
4. Nass, C., Steuer, J., Tauber, E.R.: Computers are social actors. In: CHI 1994: Proceedings of the SIGCHI Conference on Human factors in Computing Systems, pp. 72–78. ACM Press, New York (1994)
5. Sproull, L., Subramani, M., Kiesler, S., Walker, J.H., Waters, K.: When the interface is a face. *Human Computer Interaction* 11(2), 97–124 (1996)
6. Tsuduki, T., Kimura, Y.: Characteristics of media communication of college students: comparing face to face, mobile phone, mobile mail, and electronic mail. *Applied sociology studies* 42, 15–24 (2000)
7. Yamamoto, Y., Matui, T., Hiraki, K., Umeda, S., Anzai, Y.: Interaction with a computer system:-a study of factors for pleasant interaction. *Cognitive Studies* 1, 107–120 (1994)

The Effect of Physical Embodiment of an Animal Robot on Affective Prosody Recognition

Myounghoon Jeon and Infantdani A. Rayan

Georgia Institute of Technology

Atlanta, Georgia, USA

{mh.jeon, irayan}@gatech.edu

Abstract. Difficulty understanding or expressing affective prosody is a critical issue for people with autism. This study was initiated with a question, how to improve emotional communications of children with autism with technological aids. Researchers have encouraged the use of robots as new intervention tools for children with autism, but there was no study to empirically evaluate a robot compared to a traditional computer in the interventions. From these backgrounds, this study investigated the potentials of an animal robot for affective prosody recognition compared to a traditional PC simulator. For this pilot study, however, only neurotypical students participated. Participants recognized Ekman's basic emotions from both a dinosaur Robot, "Pleo" and a virtual simulator of the Pleo. The physical Pleo showed more promising recognition tendencies and was clearly favored over the virtual one. With this promising result, we may be able to leverage the other advantages of the robot in interventions for children with autism.

Keywords: Affective prosody recognition, children with autism, animal robot.

1 Introduction

Difficulty understanding or expressing affective prosody is a critical barrier to social inclusion of people with autism [1]. There has been some research on this problem [1-5], but more research is still needed for diagnosis, learning, and improvement in affective prosody. The current study explores the use of technological aids to improve the emotional communication abilities of children with autism.

Traditionally, computerized simulations of social interactions have been used as intervention tools for children with autism [e.g., 22, 23]. More recently, researchers have encouraged the use of robots [6-11], which they claim offers several advantages over traditional computerized training. Robots allow for a simplified, predictable, and reliable environment where interactions can be controlled and their complexity can be gradually increased [12]. Also, one can expect a closer rapport between children and a toy robot. A recent study [10] demonstrated that an interactive 'animal' robot facilitates social interaction more than a non-interactive robot. Using animal robots can be more beneficial because people usually have low expectation on intelligence and cognitive capabilities from them. People with autism can experience a lower workload when dealing with animals as they are often regarded as simpler than

humans. Despite these potentials with robots, there has been no study to empirically evaluate the use of a robot in the intervention compared to the use of a traditional computer.

The current study attempts to investigate the potentials of an animal robot for affective prosody recognition of neurotypical students participated. Results will serve as a baseline for future studies of children with autism.

1.1 Perception and Production of Prosody of People with Autism

Abnormal prosody has been frequently identified as one of the core features of people with autism spectrum disorder [14-16]. Because the vocal presentation of individuals with autism creates an immediate impression of oddness [17], prosody characteristics constitute critical obstacles to their social integration and vocational acceptance.

Prosody is defined as the suprasegmental properties of the speech signal that modulate and enhance its meaning [1]. Prosody functions at several levels to enable speakers to construct discourse through expressive language. Grammatical prosody is used to signal syntactic information within sentences [18]. For example, pitch contours signal the ends of utterances and denote whether they are questions or statements. Pragmatic prosody is used to carry social information beyond that conveyed by the syntax of the sentence such as intentions or the hierarchy of information. Linguistic stress can be used to highlight an element of information within a sentence as the focus of attention. Affective prosody serves more global functions including changes in register used for varying social functions and involves speaker's general emotional state [19].

There has been relatively little research on the ability of speakers with autism to perceive and understand prosodic cues [1]. Moreover, recent reviews [4, 5] have pointed to several methodological difficulties in previous studies, including small sample sizes, absence of normative data and contrast groups, poorly defined prosodic categories, and the use of subjective ratings rather than objective measures.

Previous research demonstrated that parents have more difficulty identifying the emotional content of prespeech vocalization in their children with autism than do parents of children with mental retardation or normal language [e.g., 20]. In addition, children with autism have difficulty matching vocally expressed affect to facial expressions or to emotion words (e.g., "happy," "sad," "scared") [2-3].

Studies that compare the prosody recognition ability of people with autism and neurotypical individuals have shown inconsistent results. Rutherford, Baren-Cohen, and Wheelwright [21] investigated the ability of 19 adults with HFA (High-Functioning Autism) or AS (Asperger's Syndrome) to judge the affective meaning of 40 phrases. Results showed that the HFA and AS group was impaired relative to the performance of a large number of neurotypical adults, and that the impairment did not correlate with verbal or performance IQ. However, this study used some low frequency adjectives (e.g., "derogatory," "accusatory", "intrigued") and some of their stimuli might have been semantically biased towards one of the answers. In contrast, in another study [1], both the children with HFA and the typically developing children performed near the ceiling. In this study, researchers used only "excited" and "calm" register. Thus, if the stimuli had been more complex and the emotions used had been subtler, the result might have been different. In conclusion, further exploration is needed.

1.2 Use of Robots for Autism Interventions

Some people with autism prefer to communicate with and through computers because they are predictable and place some control on the otherwise chaotic social world [22]. Indeed, Hailpern [23], in his pilot study, showed the potential of using computer-generated auditory/visual stimuli as social mirrors to encourage and reinforce sound production in children with autism.

While keeping advantages of use of computers, researchers have recently attempted to use robots for children with autism to help social interaction skills [6-11]. In addition to their basic computing ability, some robots provide a sensing and detecting environment, and can log interaction data. Since robots play a role as a toy, we can expect that psychological “rapport” between children and robots may be well formed. In fact, Mukhopadhyay [24, p. 23], who has Autism Spectrum Disorders, once reported in his book, “...his first motivation was about a set of ten bowls, colorful toys.”

Scassellati [8] showed that even non-interactive commercial robot (ESRA) can enhance motivation and engagement in children with autism. In his research, the ESRA was programmed to perform a short script which included both a set of actions and an accompanying audio file that was played from speakers hidden near the robot. Thirteen participants (7 children with Autism Spectrum Disorders and 6 neurotypical children, with mean age of 3.4 years) were positioned across a table from ESRA for a period of 3-5 minutes. Children in both groups were engaged with the robot, and often spent the majority of the session touching the robot, vocalizing, and smiling at the robot.

A more recent study [6] demonstrated that an interactive robot facilitates social interaction more than a non-interactive robot in children with and without autism, ranging from 20 months to 12 years in age. Social interaction, including the number of utterance, robot interactions, button pushes, increased more in an interactive, mobile robot, which generated bubbles only when the child pressed a button, than in the robot with random generation.

Similarly, research has shown that an interactive robotic dog (AIBO) enhances social interaction more than a mechanical toy dog (Kasha), which has no ability to detect or respond to its physical or social environment [10]. In comparison to the Kasha, researchers found that children with autism (aged 5-8 years) spoke more words to the AIBO. AIBO also elicited more social behaviors such as verbal engagement (e.g., salutations, valedictions, and general conversation), reciprocal interaction (e.g., motioning with arms, hands, or fingers to give direction to the artifact), and “authentic interaction” (e.g., the speed, tone, and volume of the child’s voice is exceptionally well modulated for the circumstances, or the child’s body is in a state of repose oriented toward the artifact as a social partner). Moreover, as the children interacted more with AIBO, they engaged in fewer autistic behaviors such as rocking back and forth and flicking fingers or hands. Also, the use of animal robots can be justified from the success in the autism literature of animal-assisted therapy (e.g., a dog [25]). It might be because animals are simpler than human, so that people with autism can feel less workload. Actually, we can see Temple Grandin’s unusual empathy with animals in her book [26].

Scassellati and his colleagues [9] have been working on affective prosody production learning in children with autism using Pleo [13], a dinosaur robot. Their pilot study was conducted with 9 to 13-year-old children with High Functioning Autism. The children must use an encouraging tone of voice to have the robot move across areas of “water”, which they are told the robot fears. Children in these trials have shown increased appropriate prosodic production with the robot trials as opposed to equal sessions with a human instructor.

1.3 The Current Study and Hypotheses

The current study explores the difference in neurotypical people’s recognition of affective prosody produced by a virtual robot and by a physical robot. Based on previous studies, participants are expected to be more socially engaged with the physical Pleo than with the virtual Pleo, leading to increased accuracy in an emotion-recognition task. In addition, participants are expected to explicitly favor the physical robot over the virtual one.

2 Method

2.1 Participants

Fourteen undergraduates and graduate students at the Georgia Institute of Technology participated in the study (5 females and 9 males, mean age 21.5). All participants signed informed consent forms and provided demographic details about age and gender.

2.2 Apparatus

Physical Pleo. A Pleo [13] was used for the physical robot condition. In addition to the advantages of use of social robots mentioned above, the Pleo has more advantages. The dinosaur is a very popular toy and its appearance is very cute. The Pleo might provide children with affordance to imitate it. The Pleo can move and have gestures so that it can elicit pointing from children, which is considered as an important gesture in autism research. Also, it can generate recorded-sound and voice. The Pleo even senses and reacts. All of these functions can easily be implemented for the research purpose because its development kit is provided by the company. Across all emotional types, the Pleo walked with the same speed without any facial expressions. In each trial, the Pleo closed its eyes, wagged its tail, and uttered a sentence with the target affective prosody before concluding with another tail wag.

Virtual Pleo. The virtual Pleo condition was created using the Pleo MySkit software on a Dell laptop PC. This software includes the P3D Performance Player, intended to allow users to review performance skits on the PC before uploading them to Pleo and included controls for navigating the performance. It uses OpenGL to draw a 3D preview of the Pleo robot.



Fig. 1. Conducting experiments with physical (left) and virtual (right) Pleo

2.3 Stimuli

Speech stimuli consisted of emotion-independent sentences (e.g., “It’s time to go”), spoken by a 31-year-old male native speaker of North American English. The speaker was instructed to produce each sentence with each of Ekman’s basic emotions (anger, disgust, fear, happiness, sadness, and surprise) [27] as well as with neutral affect. This emotion condition has been frequently used in Affect Recognition domain [e.g., 28].

2.4 Design and Procedure

A within-subjects design was used for this experiment. Independent variables included the platform (the physical Pleo and the virtual Pleo) and seven emotions.

In both conditions, participants were seated in front of the Pleo and asked to observe the Pleo as it walked and uttered the sentence. The participants chose the emotion on the given sheet. They were instructed to answer independently of the previous emotion and they did not need to necessarily mark all seven emotions. Half of the participants encountered the physical Pleo first and half encountered the virtual Pleo first. The order of appearance of the each emotion was counterbalanced across participants using the Latin Square. After completing the two conditions, participants filled out the subjective questionnaires.

3 Results

3.1 Objective Results

Emotion recognition was reasonably accurate for anger, fear, and sadness, but poor for disgust, happiness, and surprise (Figure 2). Across all emotions, recognition accuracy was greater with the physical Pleo ($M = 4.07$, $SD = 1.49$) than the virtual Pleo ($M = 3.86$, $SD = 0.95$), but this result was not statistically reliable, $t(13) = 0.52$, $p > .05$. However, detailed analysis revealed an interaction between the emotion to be recognized and the presentation type. For emotions with high recognition accuracy, both conditions showed higher accuracy in anger and sadness, compared to a chance level (50%) (see Table 1 and 2). In the case of fear, it was significantly higher than the chance only in the virtual Pleo. On the other hand, for the emotions with low

recognition accuracy, with the virtual Pleo, disgust, happiness, and surprise were significantly lower than the chance; with the physical Pleo, only disgust showed significantly lower accuracy. Consequently, the physical Pleo showed more promising recognition tendencies with graceful degradation.

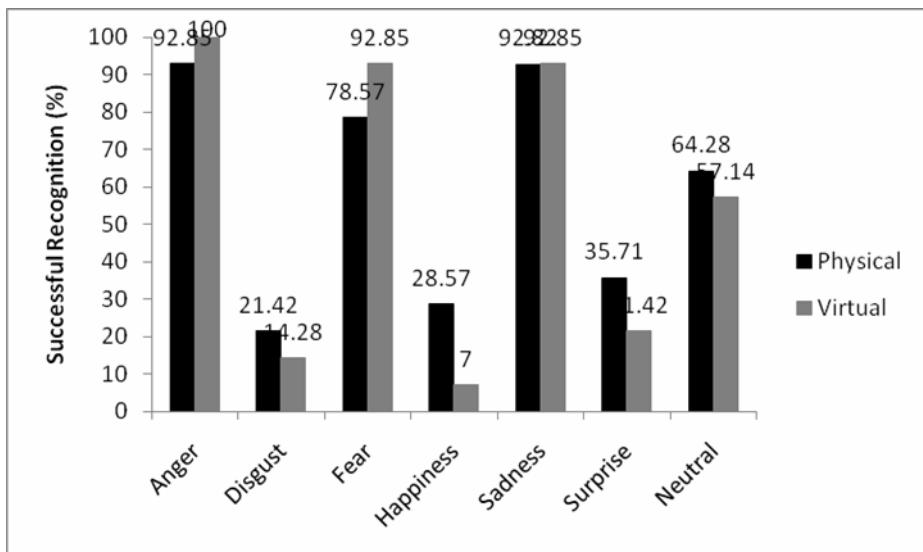


Fig. 2. Percentage of accurate responses for emotion recognition in both conditions

Table 1. One sample t-test result compared to a chance level (50%) in the physical Pleo condition

Emotion	Emotions with Positive Difference from mean of 50%		
	P-Value	Difference	t-value
Anger	<.0001	42.85	6
Sadness	<.0001	42.85	6
Fear	0.06	21.42	1.71
Neutral	0.15	14.28	1.08
Emotions with Negative Difference from mean of 50%			
	P-Value	Difference	t-value
Disgust	0.01	-28.57	-2.51
Happiness	0.06	-21.42	-1.71
Surprise	0.15	-14.28	-1.07

Table 2. One sample t-test result compared to a chance level (50%) in the virtual Pleo condition

Emotion	Emotions with Positive Difference from mean of 50%		
	P-Value	Difference	t-value
Anger	<.0001	50	infinity
Sadness	<.0001	42.85	6
Fear	<.0001	42.85	6
Neutral	0.3	7.14	0.52
Emotions with Negative Difference from mean of 50%			
	P-Value	Difference	t-value
Disgust	<0.001	-35.71	-3.67
Happiness	<0.001	-42.85	-6
Surprise	0.01	-28.57	-2.51

3.2 Subjective Results

In the subjective responses and comments, participants clearly favored the physical Pleo compared to the virtual one. Twelve out of fourteen thought that the physical presence of the Pleo was more engaging than virtual one, noting, “[the] Physical Pleo was more watchful, since I felt that we can reach out and touch Pleo,” “...immediately drew my attention, it was like seeing a toddler walk, a dog play, as opposed to a screen [the virtual Pleo],” and “People care [more] about moving-physical objects than virtual.” Also, thirteen participants preferred the physical Pleo to virtual one for the education purpose: “Virtual feels like mere recording, the physical Pleo gives the feeling that someone is talking to you,” and “It [the physical Pleo] is more fun and cool and it will attract children to spend more time with it.” Whereas some participants (7 out of 14) felt the human voice generated from the Pleo was awkward, reporting that they “wanted to be cartoon-like, it is a baby dino,” others (7 out of 14) said, “it was acceptable because [they had] already been exposed to cartoons and always wanted animals to talk,” and “it is not awkward because it is the first voice I heard attached to Pleo.”

4 Discussion

Some computer-based interventions for people with autism have led to improvements in emotion recognition [29]. Several researchers have recently attempted to improve on these techniques using interactive social robots. However, few studies have empirically evaluated the effectiveness of robots compared to desktop-computer-based interventions.

Even though the overall recognition accuracy did not differ between the two conditions, several important differences were found. When using the robotic Pleo, recognition accuracy was numerically higher, although this result did not reach statistical significance. For the lower-accuracy emotions, performance degraded more gracefully with the real Pleo. Whereas anger, fear, and sadness were relatively well recognized, disgust, happiness, and surprise were not. These results might be compared to those of children with autism and could be used for further learning of each affective prosody for various populations. Finally, participants reported a subjective preference for the robotic Pleo. This cannot be solely attributed to the novelty effect of the robots. The questionnaire showed that people are more attracted to “moving things” than something confined to a monitor. Taken together, we may be able to utilize the robot in this type of intervention with more advantages of robots than the desktop has as discussed in Introduction.

5 Future Work

This pilot study sheds light on the plausible approach to embedding affective voices on a toy robot like the Pleo. Future studies can extend this work into populations of children with High-Functioning Autism. The present study can be further developed, adding the semantic properties to the sentences. For example, children are more likely to recognize ‘happy’ from the sentence, “I really like this doll” than from “It’s time to go.” Sentences with specific semantic content may facilitate the learning of recognition of affective prosody. Additionally, the Pleo can support several gestures which could be used in a future study. A certain gesture is expected to enhance the recognition of matched emotion. Moreover, we may be able to design a more interactive study so that children with autism can use the Pleo as an agent to express their emotion or that children play and learn a ‘turn-taking’ task with the Pleo.

References

1. Paul, R., Shriberg, L., McSweeney, J., Cicchetti, D., Klin, A., Volkmar, R.: Relations between prosodic performance and communication and socialization ratings in high functioning speakers with autism spectrum disorders. *Journal of Autism and Developmental Disorders* 35, 861–869 (2005)
2. Boucher, J., Lewis, V., Collins, G.: Familiar face and voice matching and recognition in children with autism. *Journal of Child Psychology and Psychiatry* 39, 171–182 (1998)
3. Hobson, R., Ouston, J., Lee, A.: Naming emotion in faces and voices: Abilities and disabilities in autism and mental retardation. *British Journal of Developmental Psychology* 7, 237–250 (1989)
4. McCann, J., Peppe, S.: Prosody in autism spectrum disorders: A critical review. *International Journal of Language and Communication Disorders* 38, 25–350 (2003)
5. Shriberg, L.D., Paul, R., McSweeny, J.L., Klin, A., Cohen, D.J., Volkmar, F.R.: Speech and prosody characteristics of adolescents and adults with High Functioning Autism and Asperger Syndrome. *Journal of Speech, Language, and Hearing Research* 44, 1097–1115 (2001)

6. Feil-Seifer, D., Mataric, M.: Robot-assisted therapy for children with Autism Spectrum Disorders. In: IDC Proceedings – Workshop on Special Needs, June 11-13, Chicago, IL (2008)
7. Michaud, F., Theberge-Turmel, C.: Mobile robotic toys and autism. In: Dautenhahn, K. (ed.) *Socially Intelligent Agents – Creating Relationships with Computers and Robots*, pp. 125–132. Springer, Heidelberg (2002)
8. Scassellati, B.: How social robots will help us to diagnose, treat, and understand autism. In: Thrum, S., Brooics, R., Durrant-Whyte, H. (eds.) *Robotics Research. STAR*, vol. 28, pp. 552–563 (2007)
9. Scassellati, B.: Affective prosody recognition for Human-Robot Interaction. Yale University (2009)
10. Stanton, C.M., Kahn, J. P.H., Severson, R.L., Ruckert, J.H., Gill, B.T.: Robotic animals might aid in the social development of children with autism. In: ACM HRI 2008, Amsterdam, Netherlands, March, 12-15 (2008)
11. Werry, I., Dautenhahn, K., Harwin, W.: Investigating a robot as a therapy partner for children with autism. In: Proceedings of the European Conference for the Advancement of Assistive Technology (AAATE), Ljubljana, Slovenia (September 2001)
12. el Kaliouby, R., Picard, R.W., Baron-Cohen, S.: Affective computing and autism: Progress in convergence. In: Bainbridge, W.S., Roco, M.C. (eds.) *Annuals of the New York Academy of Sciences*, vol. 1093, pp. 228–248 (2006)
13. <http://pleoworld.com/Home.aspx>
14. Baltaxe, C., Simmons, J.: A comparison of language issues in high-functioning autism and related disorders with onset in children and adolescence. In: Schopler, E., Mesibov, G. (eds.) *High functioning individuals with autism*, pp. 210–225. Plenum Press, New York (1992)
15. Fay, W., Schuler, A.: Emerging language in autistic children. University Park Press, Baltimore (1980)
16. Ornitz, E., Ritvo, E.: Medical assessment. In: Ritvo, E. (ed.) *Autism: Diagnosis, current research, and management*, pp. 7–26. Spectrum Publications, New York (1976)
17. VanBourgondien, M., Woods, A.: Vocational possibilities for high functioning adults with autism. In: Schopler, E., Mesibov, G. (eds.) *High functioning individuals with autism*, pp. 27–242. Plenum Press, New York (1992)
18. Warren, P. (ed.): *Parsing and prosody: An introduction. Prosody and parsing*, pp. 1–16. Psychology Press, East Sussex (1996)
19. Bolinger, D.: *Intonation and its uses: Melody in grammar and discourse*. Stanford University Press, Stanford (1989)
20. Ricks, D.: Vocal communication in pre-verbal normal and autistic children. In: O'Connor, N. (ed.) *Language, Cognitive Deficits, and Retardation*, pp. 245–268. Butterworths, London (1975)
21. Rutherford, M.D., Baron-Cohen, S., Wieelwrigirr, S.: Reading the mind in the voice: A study with normal adults and adults with Asperger syndrome and high functioning autism. *Journal of Autism and Developmental Disorders* 32, 189–194 (2002)
22. Moore, D., McGrath, P., Thorpe, J.: Computer-aided learning for people with autism—a framework for research and development. *Innovations in Education and Training International* 37(3), 218–228 (2000)
23. Hailpern, J.: Encouraging speech and vocalization in children with autistic spectrum disorder. *Sigaccess Newsletter* 89 (September 2007)
24. Mukhopadhyay, T.: *The Mind Tree*. Riverhead Trade (2007)

25. Martin, F., Farnm, J.: Animal-assisted therapy for children with pervasive developmental disorders. *Western Journal of Nursing Research* 24(6), 657–670 (2002)
26. Grandin, T.: Thinking in pictures. Doubleday, NY (2006)
27. Ekman, P.: An argument for basic emotions. *Cognition and Emotion* 6, 169–200 (1992)
28. Bartneck, C., Reichenback, J.: Subtle emotional expressions of synthetic characters. *Int. J. Human-Computer Studies* 62, 179–192 (2005)
29. Golan, O., Baron-Cohen, S., Hill, J.J., Golan, Y.: Reading the mind in films—testing recognition of complex emotions and mental states in adults with and without autism spectrum conditions. *Social Neuroscience* 1(2), 111–123 (2006)

Older User-Computer Interaction on the Internet: How Conversational Agents Can Help

Wi-Suk Kwon¹, Veena Chattaraman¹, Soo In Shim¹,
Hanan Alnizami², and Juan Gilbert²

¹ Department of Consumer Affairs, Auburn University,
Auburn AL 36849, U.S.A

² Human-Centered Computing Division, School of Computing, Clemson University,
Clemson SC 29634, U.S.A
{kwonwis,vzc0001,szs0029}@auburn.edu,
{hanana, juan}@clemson.edu

Abstract. Using a qualitative study employing a role-playing approach with human agents, this study identifies the potential roles of conversational agents in enhancing older users' computer interactions on the Internet in e-commerce environments. Twenty-five participants aged 65 or older performed a given shopping task with a human agent playing the role of a conversational agent. The activity computer screens were video-recorded and the participant-agent conversations were audio-recorded. Through navigation path analysis as well as content analysis of the conversations, three major issues hindering older users' Internet interaction are identified: (1) a lack of prior computer knowledge, (2) a failure to locate information or buttons, and (3) confusions related to meanings of information. The navigation path analysis also suggests potential ways conversational agents may assist older users to optimize their search strategies. Implications and suggestions for future studies are discussed.

Keywords: Conversational agent, older users, Internet, interaction.

1 Introduction

Only 35% of seniors aged over 65 years use the Internet in the United States [1]. The primary cause for this underrepresentation of older users is that hardware and software interfaces have not been designed to accommodate their needs [2].

Conversational agents (CAs) are animated embodiments in computer-mediated environments that respond to users through verbal and non-verbal communication [3]. CAs facilitate 'intelligent' retrieval of information based on users' individual needs. Applying findings from research on pedagogical agents [4], we propose that CAs have a significant potential to improve older users' interactions in computer-mediated environments. The purpose of this study was to identify potential roles a CA can play to enhance older users' Internet interactions in consumer environments, through a qualitative study employing a role-playing approach with human agents. The researchers assume that real interactions with humans can be used to model interactions with CAs.

2 Methods

Participants (12 men, 13 women; ages of 65 years or older) were recruited among members of a lifelong learning center at a Southeastern university. Participants met in groups of four to eight. The study was conducted in a conference room equipped with laptop computers, external mice, and high-speed wireless Internet connections. After answering general questions on computer and Internet experience, participants performed a shopping task on a given e-commerce site (Amazon.com) to shop for a product (a pair of sneakers or a flat-screen TV). Amazon.com was chosen because it represented the typical level of complexity found on e-commerce sites, and it is one of the most used sites by U.S. consumers. A human assistant who played the role of a CA sat next to each participant to answer questions or to actively guide participants through the shopping task. The shopping task computer screens were video-recorded using video capture software, Debut. In addition, the participant-assistant conversations were recorded through audio recorders.

The audio-recordings of the participant-assistant conversations were transcribed, and all meaningful segments of the verbatim were subjected to content analysis to discover common themes. The screen-recordings of the shopping tasks were used to identify the navigation paths via which participants completed their shopping task (see Fig. 1 for an example path).

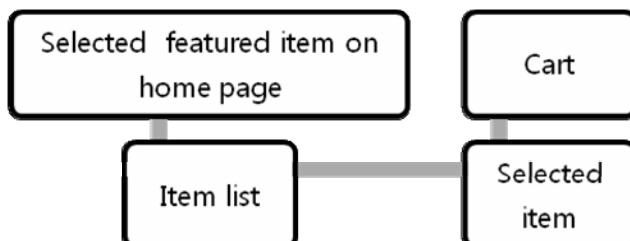


Fig. 1. Example of a navigation path adopted by a participant to buy a TV

3 Findings and Discussion

The participant-assistant conversation verbatim content analysis revealed three main types of difficulties participants experienced for which CA assistance could be employed (see Table 1). First, *lack of prior computer knowledge* was an obstacle facing eight of the 25 participants. This was evident when participants could not locate an Internet browser, figure out how to launch www.Amazon.com on the browser, or find keys on the keyboard, and when they used wrong mouse clicks (e.g., left/right clicks, single/double clicks). However, even among those who successfully accessed the site, a number of problems related to information overload and complexity arose. Older adults experience declining cognitive abilities such as reduced perceptual speed, visualization ability, working memory, and attention, which can lead to slower information search and retrieval performance [5]. Thus, older users may be more overwhelmed by the amount and complexity of information presented on a Website.

Table 1. Problems experienced by older adults in web-interactions needing agent assistance

Problems needing assistant help	Number of participants	Example participant inquiry
Lack of prior computer knowledge	8	<ul style="list-style-type: none"> • Yeah, I'm having a hard time finding Internet Explorer. • Where's the enter key? • Oh, you have to double click? On mine I don't have to. • Where is Amazon? • Did it go through? How did you know [that it went through]?
Failure to locate buttons/information	11	<ul style="list-style-type: none"> • Select color... Ok. Alright. I don't see options to... color? • Is there a way we can cluster the 19 inch TVs together to make a price comparison? • OK, where are the "32 results" [search results]. I want to see the results come out. • I don't see where it says to buy it... I wish we can see all buying option • What? Where is shopping cart? • Let's see... Where would I... I'm looking because I'm not familiar with it.
Confusion about information meanings	4	<ul style="list-style-type: none"> • "Hot dog treats" [advertisement]? Why did it put all this stuff? • Espresso [color]? What's espresso? • Seven and a half 2A... Wait a minute. What does this mean?

The content analysis also showed that 11 participants sought for their assistants' help for a *failure to locate information or buttons* that were presented on the screen such as a price comparison button, product category menu, and buttons to put selected items into the shopping cart. In addition, four participants complained about their *confusions related to meanings of information* presented on the screen and requested their assistants' help. For example, participants were confused by peripheral information such as advertisements. Other participants were not able to comprehend some of the retailer terminology such as product color names, dynamic pricing systems, and product sizing systems.

The navigation path analysis revealed how CAs can optimize the search strategies of older users. For example, participants used broad search results (instead of employing specific search terms) or navigated a long list of items to find the product of their choice. They were unaware of options available on the side menus (e.g., brand and other

product criteria) to narrow their search. These broad searches led to information overload and reduced success in finding products that met the participants' criteria.

By providing real-time responses to users' inquiries and/or by actively seeking to understand users' problems and needs and proactively guiding them to reach their goal, CAs may be able to relieve some of the anxieties older users experience in dealing with the Website information overload and complexity. CAs can also help older users narrow their search strategies and lead to greater efficiency in finding the right products. Future research is needed to expand our understanding of how intelligent agent interfaces can be designed to benefit older users in web-based consumer environments. Specifically, research is recommended on the optimal levels of CA interactions (e.g., responsive CAs vs. proactive CAs, functional vs. social CAs) depending on diverse needs of older users.

Acknowledgment and Disclaimer. This material is based in part upon work supported by the National Science Foundation under Grant Number IIS-0955763. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

1. Pew Internet,
http://www.pewinternet.org/trends/User_Demo_7.22.08.htm
2. Bucar, A., Kwon, S.: Computer Hardware and Software Interfaces: Why the Elderly Are Under-represented as Computer Users. *Cyberpsychol. Behav.* 2, 535–543 (1999)
3. Cassell, J., Sullivan, J., Prevost, S., Churchill, E.F.: *Embodied Conversational Agents*. MIT Press, Cambridge (2000)
4. Baylor, A.L.: Intelligent Agents as Cognitive Tools for Education. *Educ. Technol.* 39, 36–40 (1999)
5. Czaja, S.J., Sharit, J., Ownby, D., Roth, D., Nair, S.N.: Examining Age Differences in Performance of a Complex Information Search and Retrieval Task. *Psychol. Aging* 16, 564–579 (2001)

An Avatar-Based Help System for Web-Portals

Helmut Lang¹, Christian Mosch², Bastian Boegel³,
David Michel Benoit⁴, and Wolfgang Minker¹

¹ Institute of Information Technology, Ulm Univ., 89081 Ulm, Germany

² Communication and Information Center, Ulm Univ., 89069 Ulm, Germany

³ Institute of Information Resource Management, Ulm Univ., 89069 Ulm, Germany

⁴ Institute of Theoretical Chemistry, Ulm Univ., 89069 Ulm, Germany

{helmut.lang, christian.mosch, bastian.boegel,
david.benoit, wolfgang.minker}@uni-ulm.de

Abstract. In this paper we present an avatar-based help system for web-portals that should provide various kinds of user assistance. Along with helping users on individual elements of a web page, it is also capable to offer step-by-step guidance supporting users to complete specific tasks. Furthermore users can input free text questions in order to get additional information on related topics. Thus the avatar features a single point of reference, when the user feels the need for assistance. Different to typical systems based on dedicated help sections consisting of standalone HTML pages, help is instantly available and displayed directly at the element the user is currently working on.

Keywords: HCI; Computer Assisted Learning; Grid Computing.

1 Introduction

The recent tendency in personal computing to shift content from a local machine to the cloud has led to a new kind of interfaces. Applications running on a local machine are subsequently replaced by programs accessed by web browsers. The success of netbooks and the development of Google's chrome OS are reflecting these changes. Nevertheless web based systems often do not feature equal ease of use, as programs run locally. This is mostly due to deficiencies in current web browsers along with immature standards.

Web based help systems nowadays usually consist of a set of HTML pages that force users to switch their focus of attention from the page they are currently editing towards this dedicated help pages. Another downside of an HTML based approach is that users, especially those new to the given field, have the feeling to get lost. [16]

A web-portal currently developed in the context of the bwGRID project [2] should overcome this problem. The portal allows users to access grid computing resources from within their web browsers. Thus we decided to implement an avatar-based help system that offers the information needed by users for the completion of certain tasks right on the web page they are currently editing.

In addition to task specific guidance the avatar shall be able to provide further information on related concepts.

The following sections we will first introduce the architecture of the web-portal used at our site along with the implications for a help system caused by the architectural approach. Then we define the demands for the help system according to a typical use case of the portal. Finally we describe, how the avatar-based system meets those demands and provide some details on its technical realization.

2 Portal Architecture

Grid computing usually requires massive usage of command line tools. Those hard to learn interfaces increase the barrier for using the grid especially for inexperienced users [1]. The objective of the bwGRiD portal project [3] is to make grid computing more appealing to novice users by offering access to grid computing resources by means of a web-portal.

In order to provide interfaces for different applications each project partner is developing portal based applications fitted for specific target audiences. The portal developed in Ulm addresses the needs of theoretical chemists. Hence it features a graphical molecule editor allowing users to prepare input files for Gaussian [13] and NWChem [12] and enables “point and click” driven submission of the generated jobs to the grid.

Fig. 1 depicts an architectural overview of the portal in use at our site. Client communication is handled by a tomcat servlet container in combination with GridSphere [9]. GridSphere is an implementation of the JSR 168 Portlet Specification [10]. Different to servlets that return web pages as a whole, portlets are only responsible to render fractions of pages. When the portlet container (GridSphere) receives a client request, it determines all portlets included in the requested page, triggers them and integrates the returned output into a response.

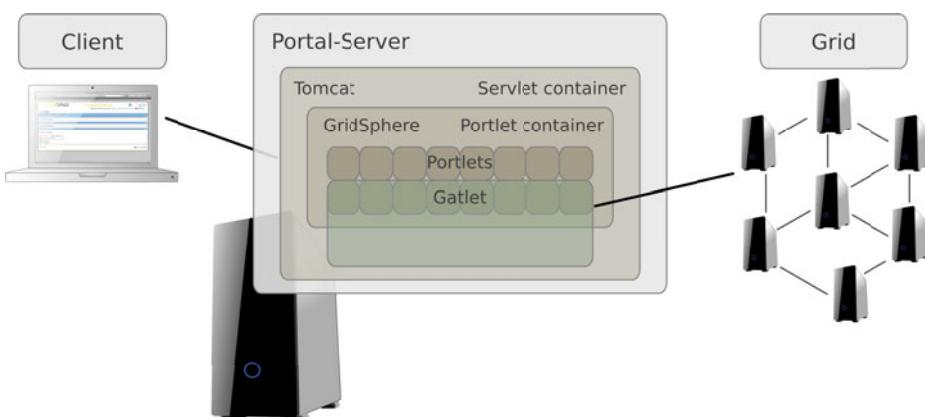


Fig. 1. Architectural overview of the portal in use at our site

The modular nature of portlets perfectly fits the idea of different sites contributing different modules for various applications. However this modularity makes developing an integrated help system for such a portal quite time and resource consuming. Also the help system has to be designed in modular way and each portlet has to provide the information needed in a self-contained manner. The system then has to combine the information available for every portlet into a unified user experience.

The portal's link to the grid is established by Gatlet [6]. Gatlet offers a set of generic portlets needed for grid access in general (e. g. to administer MyProxy credentials or to access files stored on the grid) along with an API for interaction with grid resources. Currently Gatlet features support for the Globus [8] grid-middleware but support for UNICORE [17] and gLite [7] is planned.

3 Use Cases and Requirements

A typical workflow for a chemist submitting a job to the grid would involve the following steps:

1. Login to the portal using a certificate stored in the web browser's keystore.
2. Navigate to a page containing the MyProxy portlet provided by Gatlet.
3. Use the MyProxy portlet to retrieve a proxy credential.
4. Navigate to the page containing the chemistry portlet.
5. Use the graphical editor to create a molecule.
6. Choose if you want to create an NWChem or Gaussian job and edit a set of parameters.
7. The input file generated by the portal can be edited, in case the user wants to make manual adaptations.
8. Choose the resource the job should be executed at.
9. Navigate to a page containing the job monitoring portlet.
10. Observe the status of the job until it is finished.
11. Retrieve and examine the results.

Analyzing this workflow points out some important features that an adequate help system should provide:

- There is a well-defined sequence of steps that have to be processed to complete the given task. Thus for novice users a step-by-step guide visiting all required portlets and assisting the user in taking decisions (e. g. if creating a Gaussian job is preferable over an NWChem job) is really helpful. Due to the modularity of the portal, the portlets involved during the step-by-step tutorial might be spread over different pages. Thus the system has to be aware of the portal layout and the guide has to span multiple pages.
- Users already familiar with the portal might only need help on specific parts of the interface. In this case the system should be able to provide information on individual elements.

- Interested users are also confronted with new concepts, like “credentials”. Even though understanding the relevance of proxy credentials is not essential to use the portal, users might be interested to get to know more about grid computing in general. Hence the help system should also be able to give information on continuative topics related to grid computing and chemistry. This feature is of particular relevance, since the target audience of the portal are people new to grid computing, like students.

We came to the conclusion, that these features can be implemented with an avatar-based system quite elegantly and that such a system would be superior to other approaches. The advantages of such an avatar-based framework will be discussed in the following section.

4 Features

When the avatar is idle he is displayed as a small icon along with an “Assist me” button (see Fig. 2). In order to keep the avatar from interfering with other elements on the page the user can drag him to different locations.

While trying to exploit the well known “persona effect” [11] we also wanted to keep the effect of reduced performance, achieved when the user feels observed by an animated agent [14], as low as possible. Hence the avatar does not exhibit any idle animations.

To get help from the avatar the user has to click on “Assist me”. After this the actual content of the page is shaded, a short introductory message is displayed in

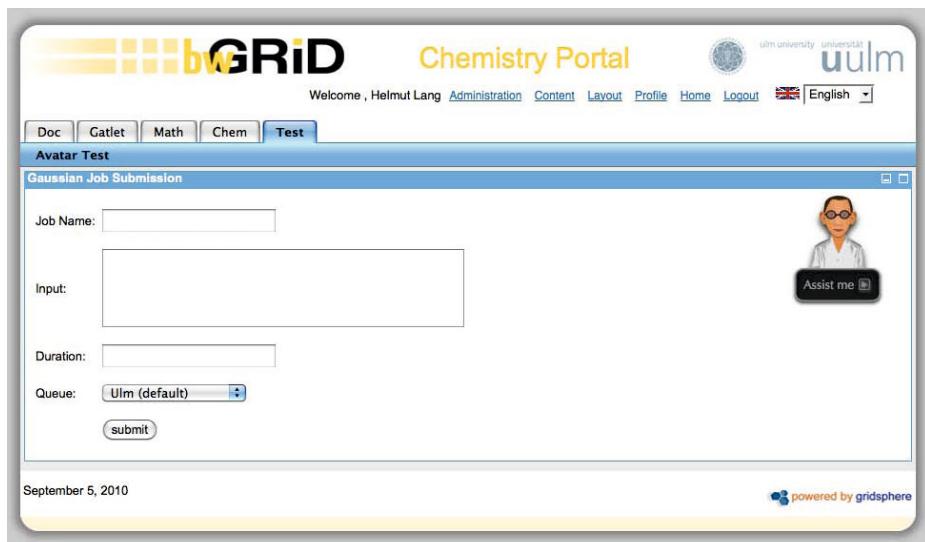


Fig. 2. The avatar in idle state

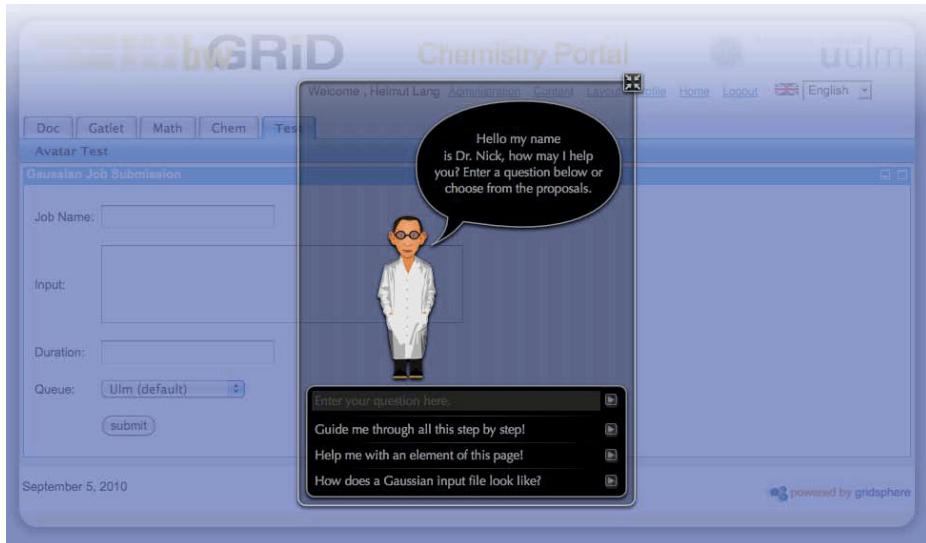


Fig. 3. The avatar after clicking “Assist me”

the avatars speech balloon and different options for further actions are presented in an interactive area below the avatar (see Fig. 3).

4.1 Element Explanation and Input Verification

The basic functionality of the avatar is to explain elements on the page and to verify the user input. To get help on an element the user chooses “Help me with an element of this page” at the greeting dialogue (see Fig. 3). Then all elements the avatar is able to explain are distinctively marked. When the user clicks on one of the marked components the avatar points at it and provides additional information (see Fig. 4).

This behavior implements the second requirement from the previous section, by providing information on elements chosen by the user. In contrast to tooltip or balloon help systems, our approach allows users to identify fields with additional information at first sight, without trying to keep the mouse as still as possible over an element. During explanation, the current input field remains editable. As a side effect text displayed in the speech balloon, may be copied and pasted. This is especially useful to provide templates for input fields of higher complexity.

Displaying the information along with the actual content reduces the cognitive load on the user, compared to dedicated help pages. 4

Another advantage of the avatar is the capability to compensate for deficiencies of the underlying interface. In the example of Fig. 4, the user would normally have to enter a duration from 120 to 604 800 in seconds into an input field. The avatar superimposes this field with separate spin boxes for days, hours

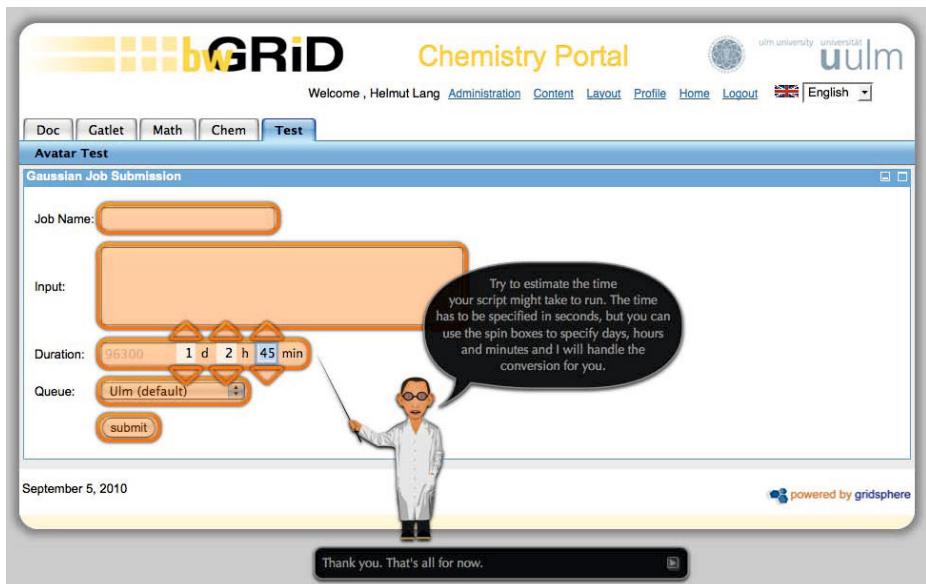


Fig. 4. The avatar assisting a user with entering a duration

and minutes, thus relieving the user from the burden of doing excessive calculations manually. Furthermore it is also possible to define minimum and maximum values and the avatar takes care that those limits are kept.

Aside from validating durations, it is possible to define a regular expression for input verification. In case of user input violating this expression, the avatar clarifies which input is actually allowed.

4.2 Step-by-Step Guidance

The avatar is able to assist users performing certain tasks utilizing step-by-step tutorials. Thus satisfying the first requirement from the previous section.

In the step-by-step scenario the avatar directs the user to the pages containing portlets that need to be visited during the step-by-step tutorial by pointing at the corresponding items in the navigational menu. On these pages all elements that need to be edited are explained sequentially. That way we have realized a typical “worked example” [15] kind of learning.

4.3 Continuative Information

The input field in Fig. 3 is used to enter questions covering continuative topics. If an answer to the entered question is found it can be presented to the user as a kind of slide show, where the avatar makes remarks on certain elements. Fig. 5 shows how the avatar explains the constituents of Gaussian input file. Next to this rather sophisticated way of presentation, it is also possible to just display some text in the speech balloon, depending on the complexity of the answer.

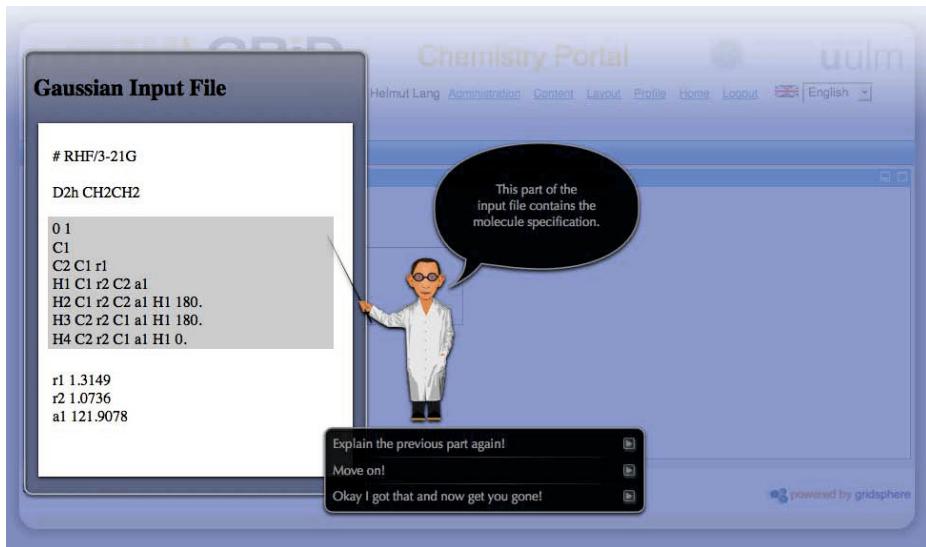


Fig. 5. The avatar explaining the constituents of a Gaussian input file

```
<?xml version="1.0" encoding="UTF-8"?>
<av:answer xmlns:av="http://uni-ulm.de/bwgrid/portal/avatar"
  xmlns="http://www.w3.org/1999/xhtml"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://uni-ulm.de/bwgrid/portal/avatar
    ../schema/answer.xsd ">

<av:complex>
  <av:slide title="Gaussian Input File">
    <av:content>
      # RHF/3-21G<br/><br/>
      D2h CH2CH2<br/><br/>
      <div id="molSpec">
        0 1<br/>
        C1<br/>
        C2 C1 r1<br/>
        H1 C1 r2 C2 a1<br/>
        H2 C1 r2 C2 a1 H1 180.<br/>
        H3 C2 r2 C1 a1 H1 180.<br/>
        H4 C2 r2 C1 a1 H1 0.<br/>
      </div>
      <br/>
      r1  1.3149<br/>
      r2  1.0736<br/>
      a1 121.9078<br/>
      <br/>
      <br/>
    </av:content>

    <av:explainPoint ref="molSpec" pos="EAST">
      <av:text>This part of the input file contains the molecule specification</av:text>
      <av:forward>Explain the previous part again!</av:forward>
      <av:backward>Move on!</av:backward>
    </av:explainPoint>
  </av:slide>
</av:complex>
</av:answer>
```

Fig. 6. A reply (abbreviated for clarity) as returned for a question like “What does a Gaussian input file look like?”

Furthermore the system suggests context dependent questions, displayed as options in the greeting dialogue (see Fig. 3). That way users get an idea on how to formulate questions for the avatar and which kind of questions the avatar might be able to answer. Logging the questions asked on certain pages can be used to determine frequently asked questions and make appropriate suggestions.

5 Technical Realization

5.1 General

The client side of the avatar is completely implemented in JavaScript. We have made a significant effort in keeping it compatible to a broad range of contemporary browsers. By using JavaScript only the avatar is completely independent of plug-ins and additional software.

5.2 Element Explanation and Input Verification

The text prompted by the avatar, when the user asks for help on certain elements is embedded into the page by means of hidden divs. To generate those hidden divs a tag library was implemented that allows extending the JSP-pages used to render the portlet. The following code example shows how the component helping the user to edit durations is added to a page:

```
<%@ taglib uri="http://uni-ulm.de/bwgrid/avatar/" prefix="av" %>
<input type="text" size="20" id="duration"/>
<av:compDesc key="durHelp" ref="id:duration" type="DurationHelper"
  constraints="min=60, max=604800" next="queueDesc"/>
```

The first part imports the required tag library, the second defines the actual input field and the remainder is responsible for generating the content relevant to the avatar. The `key` attribute in the last tag refers to a value in a Java properties file. This allows to define text for multiple languages.

Next to the `DurationHelper` type a `RegexpHelper` allows specifying a regular expression as constraints for the accompanying input field.

5.3 Step-by-Step Guidance

Within portlets, steps are defined specifying a `next` attribute within a `compDesc` Element. To allow step-by-step guides to span multiple portlets, developers have to create an additional `avatar.xml` file, containing information about which portlet to visit, after the user is done with the respective portlet. This file is added to the list of “watched resources” in the tomcat server. Each time the file is changed a listener is triggered and the component holding the meta information about the portal is updated.

In case the user asks for a step-by-step guide, an asynchronous request is sent to a servlet on the portal server. According to the stored meta information a list of portlets, along with the pages they can be found at, is dynamically created and returned to the client. This list is afterwards kept in the browser’s session storage and sequentially processed by the avatar.

5.4 Question Answering

When the user enters a question into the input field of the greetings dialogue or chooses from one of the suggestions a request to the server is triggered. To find an appropriate answer the utterance is searched for matching keywords. The following reply consists of an xml-document constituting the answer. The system distinguishes between simple and complex answers. The suitable way to present the content is chosen by the content editor. A simplified example for the structure of a reply constituting a complex answer can be seen in Fig. 6. The content of this reply would be presented by the avatar as can be seen in Fig. 5.

5.5 GridSphere Dependencies

The approach chosen for the element explanation functionality and the answering of questions does not exhibit any GridSphere dependencies and thus can be easily adapted to other portals relying on the Java servlet technology.

In contrast to this step-by-step guides are highly GridSphere dependent. This is owed to the fact that there has to be a mapping between portlets and pages they appear on. GridSphere stores this information in a dedicated file that is scanned each time the user issues a request for a step-by-step guide. Since the method layout information is stored differs between portlet containers, adapting the step-by-step functionality to other systems would require additional work.

6 Conclusion

We propose an integrated help system that is able to assist users, without having to switch their focus of attention from the page they are currently working on. Thus help is instantly available, where it is needed, when it is needed and from a unified resource. The avatar smoothly integrates several different modes of support. He is capable of providing users with step-by-step guidance, information on continuative questions and help on individual elements. Due to this variety of features the avatar is beneficial for novice as well as advanced users.

A distinctive feature of our system is the capability to automatically integrate individual portlets on different pages into a coherent step-by-step tutorial.

Apart from this step-by-step functionality that is dependent on GridSphere – the portlet container in use at our site – all components of the system can easily be ported to portals relying on the Java servlet technology. To verify the portability we installed the avatar on a portal running Liferay [5] and only minimal adaptations where needed. Consequently the system is applicable for a wide range of application.

Future work on the project will involve user studies to verify our theses on the usability of the avatar and to draw conclusions for improvements.

Acknowledgments. The bwGRiD portal project is funded by the Ministry of Science, Research and the Arts Baden-Wuerttemberg.

References

1. Altmann, A.: Direkte Manipulation: Empirische Befunde zum Einfluß der Benutzeroberfläche auf die Erlernbarkeit von Textsystemen. Zeitschrift für Arbeits- und Organisationspsychologie 31(3), 108–114 (1987)
2. bwGRiD Project, <http://www.bw-grid.de/>
3. bwGRiD Portal Project, <http://www.bw-grid.de/portal/>
4. Chandler, P., Sweller, J.: Cognitive Load Theory and the Format of Instruction. Cognition and Instruction 8(4), 293–332 (1991)
5. Enterprise Open Source Portal and Collaboration Software – Liferay.com, <http://www.liferay.com/>
6. Gatlet Project Page, <http://gatlet.scc.kit.edu/index.php>
7. gLite - Lightweight Middleware for Grid Computing, <http://glite.cern.ch/>
8. Globus Alliance, <http://www.globus.org/>
9. GridSphere Portal Framework, <http://www.gridsphere.org>
10. JSR-000168 Portlet Specification (Final Release), <http://jcp.org/aboutJava/communityprocess/final/jsr168/index.html>
11. Lester, J.C., Converse, S.A., Kahler, S.E., Barlow, S.T., Stone, B.A., Bhogal, R.S.: The Persona Effect: Affective Impact of Animated Pedagogical Agents. In: Proc. CHI 1997, pp. 359–366 (1997)
12. NWChem High-Performance Computational Chemistry Software, <http://www.nwchem-sw.org>
13. Official Gaussian Website, <http://www.gaussian.com/>
14. Rickenberg, R., Reeves, B.: The Effects of Animated Characters on Anxiety, Task Performance, and Evaluations on User Interfaces. In: Proc. CHI 2000, pp. 49–56 (2000)
15. Sweller, J., Cooper, G.A.: The Use of Worked Examples as a Substitute for Problem Solving in Learning Algebra. Cognition and Instruction 2(1), 59–89 (1985)
16. Theng, Y.L., Thimbleby, H.: Addressing Design and Usability Issues in Hypertext and on the World Wide Web by Re-Examining the “Lost in Hyperspace” Problem. Journal of Universal Computer Science 4(11), 839–855 (1998)
17. UNICORE (Uniform Interface to Computing Resources), <http://www.unicore.eu/>

mediRobbi: An Interactive Companion for Pediatric Patients during Hospital Visit

Szu-Chia Lu, Nicole Blackwell, and Ellen Yi-Luen Do

Georgia Institute of Technology, GVU Center

Atlanta, GA 30318, USA

{sj.today, nblackwell13, ellendo}@gatech.edu

Abstract. Young children often feel terribly anxious while visiting a doctor. We designed mediRobbi, an interactive robotic companion, to help pediatric patients feel more relaxed and comfortable in hospital visits. mediRobbi can guide and accompany the pediatric patients through their medical procedures. The sensors and servomotors enable mediRobbi to respond to its environmental inputs and the reactions from young children as well. The ultimate goal of this study is to transform an intimidating medical situation into a joyful adventure game for the pediatric patients.

Keywords: Pediatric care, robotics, Children behavior.

1 Introduction

While some young children are confident about visiting the pediatrician or dentist, others are filled with dread and anxiety at the thought of the experience. This fear-filled event can be unpleasant for parents and children alike. The visit can quickly become a battle before, during or even after the exam. The reason young children experience this anxiety is the fear of the unknown, which is often more frightening than the reality. Under these circumstances, a tool to put the child's mind at ease would be greatly appreciated. An explanation, including what to expect from the doctor or nurse, would be beneficial in helping to mentally prepare the child for the process [1].

However, an explanation alone could not eliminate anxiety once the medical procedure begins. If an uncomfortable step needs to be performed (a simple throat swab, a scheduled vaccination, a weight/height recording, or a blood pressure check), the child needs to feel safe and secure about the impending proceedings.

For this reason, having a comforting object with them during the doctor's visit can provide further relief. Children can usually benefit from having a comforting object with them during their physical exams [2][3]. These friendly objects can help a child remain calm. Young children find comfort in a stuffed animal or a security blanket; older children may choose a favorite article of clothing or a book. These comforting objects play the role of a sharing companion, a consolatory pet or a brave friend that helps the child remain confident while redirecting the attention to something else.

An early report described the benefits of Animal-Assisted Therapy (AAT) [4], in which the animal companion would serve as a communication link, and provide the

child, even the one with autism, with a sense of security in the therapy setting and perceived shortened therapy process. For example, Boris Levinson, M.D., a Canadian child psychiatrist, included his dog in therapy sessions [5]. However, not all pediatric hospitals can afford to have animal companion for each child during their hospital visit. Therefore, we are interested in developing a robotic pet to serve this companion function.

In this project, we utilized the Lego Mindstorms' building components to design and program an interactive robotic pet called mediRobbi. Equipped with various types of sensors, mediRobbi can detect light levels, distances of other objects in front of the robot, sounds from the surroundings, and touches from people or other objects. These sensors enable mediRobbi to receive signals from the pediatric patient, doctor, or even the environment. Signals such as colors or symbols on a medicine label, distance of the medical equipment, sound from the young patients, or behaviors of the doctor can all be perceived by the mediRobbi robot. Most importantly, the mediRobbi can generate motor, visual, or audible feedback in response to those inputs through its servomotors. Through the companionship of mediRobbi, the ultimate goal of this project is to help the pediatric patients transform their intimidating medical experiences into a joyful game play.

2 Related Work

2.1 Comforting Objects during Hospital Visit

Examples of comforting objects in hospitals are abundant. For example, the Sons of the American Legion in Hays initiated the "comforting kit" program to ease the intimidating experience of the pediatric patients who spend the night at Hays Medical Center [6]. They send out the comforting kit called "a Josh and Friends kit". The package includes an adorable plush puppy "Josh" and a book, named "I'll Be OK" [7]. Both the puppet and the storybook were packaged in a doghouse that is ready for those young patients who are visiting the hospital for the first time. The director of Women and Children Nursing Services at Hays Med, Celeste Gray, stated that this comforting kit has helped make the hospital more of an adventure for the child and possibly relieve some of the stress.

Meanwhile, another institute, the Emerson Hospital, in Massachusetts also distributed bags of toys and books, called Coping Kits, to comfort their pediatric patients and to support the parents during visits to their hospital's Emergency Department [8]. The idea of this Coping Kits was developed by the hospital's Pediatric Intervention Team, including a Child Life Specialist, a registered nurse and the team's nurse coordinator. The content for each Coping kits varies for children by age and color coded by gender. Every kit contains five tools for five different coping behaviors for children: 1) distract from the stressful exam process, 2) encourage relaxation and deep breathing, 3) provide comfort, 4) require focusing on an object, and 5) encourage communication and self-expression.

The above examples of hospital practices demonstrate the importance of a companion or a comforting object for young patients to alleviate the stress during the medical process and to bridge the communication between the medical staff and the

children. However, these coping and comforting kits cannot produce real-time responsive support. The toys or storybooks can be beneficial to prepare the pediatric patients for their hospital visits or serve as great distractions for children to focus their attention on something other than the dreadful medical procedures. However, they cannot respond to those young patients' behaviors or reactions. Young children might get bored very easily with unresponsive objects. Therefore, we designed mediRobbi to be interactive and responsive.

2.2 Human-Companion Animal Interaction

Animal-Assisted Therapy (AAT) is another therapeutic aspect about interaction between the young children and their pets to provide strong emotional bonds that serves as a source of support to the pet owners [4][5][9]. We believe our young patients can benefit from this kind of support. Several psychology studies validated that pets could contribute to the development of 1) a child's basic **sense of trust** through the pet's constancy, security, reliability, love and affection; 2) a **sense of autonomy** through the pet's serving as an active playmate and encouraging patience and self-control; 3) a **sense of industry** through the pet's trainability and response to the child's basic commands; and 4) a **sense of identity** through the pet's serving as a companion, and providing social and emotional support [10][11][12]. These four senses are quite important either for just a short-term one-time hospital visit or a long-term chronic treatment for pediatric patients.

These Animal-Assisted Therapy studies reported that responsive pets served as an excellent communication link that provides the young patients with a sense of security in the hospital setting. These responsive companions quickened the medical process as well.

3 Initial Prototype of mediRobbi Using Lego Mindstorms

For a proof-of-concept prototype of this interactive companion, Lego Mindstorms NXT was used to build this interactive robot. Lego Mindstorms is a line of programmable robotics construction toy, manufactured by the Lego Group [13]. It comes in a package containing many pieces including sensors and cables. These input sensors including a touch sensor that enables the robot to feel and react to its environment, a sound sensor so that the robot can hear and react to sound, a light sensor to detect light and color for the robot, an ultrasonic sensor providing the robot the ability to see, measure distance to an object, and react to movement, and, at last, three interactive servo motors that ensure the robot movement with precision. With the help of these sensors, the robot will be able to receive signals from the pediatric patient, doctor, or even the environment.

We followed the building instruction in Lego's Robo Center and built a rapid prototype of an interactive pet by using the input sensors and the building bricks (See Figure 1). It was equipped with all five types of sensors, a touch sensor, a sound sensor, a light sensor, an ultrasonic sensor, a light sensor and three servomotors. The robot even has hands to hug or grab small objects.

Then, the robot was programmed into sets of reaction orders through the Lego Programming Palette (See Figure 2). The programming palette contains all of the programming blocks a creator will need to create programs. Each programming block determines how the robot will act or react. By combining blocks in sequence, a creator can quickly create programs that will make the robot come to life.



Fig. 1. In top-right, the middle model is the interactive pet we built. Around the interactive pet are the input sensors used in the model.

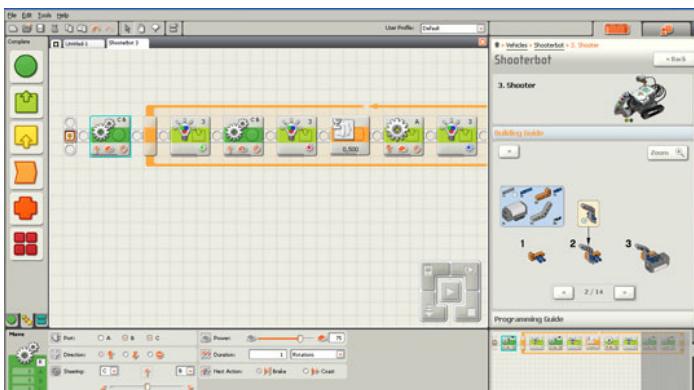


Fig. 2. The Lego Mindstorms programming palette. Using the drag-and-drop command block to design the interaction for the robot.

There are five interactions sets for the pet robot. These interaction sets were designed to play the game Stop-and-Go, which is a board game explaining allergy related information to the children.

The first interaction set is activation. Pushing the stick to press down the touch sensor, which is underneath the head of the robot, can activate the robot. After putting

the robot on the map and activating it, the robot will follow the order of next interaction set to follow the black route on the map. This activation feature is critical in the play of the adventure board game, and in the next phase, in leading the pediatric patients through their examination as well.

The second and the third interaction sets were accomplished by using the light sensor equipped facing down to read the black track or red spot on the white map. The light sensor can detect the degree of brightness, and adjust the motor movement accordingly.

The fourth design of interaction is that the robot will reactivate after stopping at a red spot by a loud sound or voice, such as clapping or shouting out. This interaction can be triggered by the sound detected by its sound sensor.

The last interaction feature is visual and audible feedback from the robot. When the robot reaches the home station on the map or achieves a task goal, it will trigger the hooray sound effect download to the robot, and a simple dancing movement of clapping its hands.

4 Experiment Design

In order to examine whether the children would enjoy the company of their new friend during a stressful situation of a hospital visit, we are conducting a paired interview of children and their parents and a field observation study of children in the hospital.

4.1 Paired Interview

The paired interview involves young children in the age range of 3 to 7 and their parents. The interview includes questions about the experience of hospital visit of young children and how their parents prepare them mentally or physically before or during the hospital visits.

A simple questionnaire and consent form will be conducted before the experiment. The questionnaire will collect the following information:

1. *About the parents.* Age range of the parents, how many children (in the age range of 3-7) within the family, frequency of hospital visit, approximate duration in one regular hospital visit (including the waiting time), how they prepare the child before/during the hospital visit, and specific reaction (if any) during the hospital visit.
2. *About the child.* Age, gender, how they feel when visiting a doctor and why, what they particularly like/hate during the hospital visit, would they like to have a companion during the hospital visit?

These data were collected through interview in person or via phone, and stored in electronic format in a secured online database.

4.2 Field Observation of Children Behavior

The field observation was conducted in a pseudo hospital environment. We utilized the empty lab space into a pediatric examine room and a children waiting/playing area. There are three phases for the field observation: 1st phase of Introduction, 2nd

phase of Familiarization and 3rd phase of Real Examination. Each phase has its purpose for young participants to know the companion, mediRobbi, to get familiar how mediRobbi would react to their behavior or orders, and then to start the exam process after they feel comfortable being with their companion.

First Phase - Stage of Introduction. The purpose of this phase is to present the mediRobbi to our young participants and to start building the trusting connection between them and the companion. In the Introduction phase, the young participants were asked to choose the name and gender of their companion from an interactive webpage (see Figure 3). Then, the young patients decided how they want to decorate or dress their companion by choosing the clothing pattern from the same website

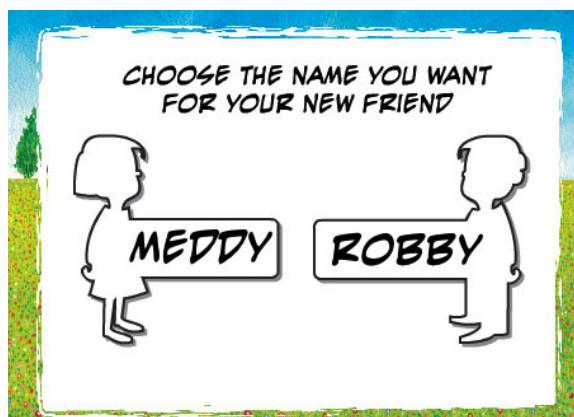


Fig. 3. Snapshot of the website for the children to choose name and gender for their new companion



Fig. 4. This is the snapshot of the second step for children to choose one of decoration pattern for their new companion

(see Figure 4). After the choosing step, the premade origami shirt or dresses (see Figure 5) was attached to mediRobbi and present the personalized companion to the young participant.



Fig. 5. An origami dress is attached to the interactive companion

Second Phase – Stage of Familiarization. The second phase is for the young patients to familiarize themselves with their new playmate by running their mediRobbi on an educational allergy board game, called *Stop-and-Go*, a 36"x48" carpet-sized map. The main purpose of this game is to increase the sense of ownership of the young participants and to build the trusting relationship between the young participant and the mediRobbi through the leading-following process. During the gameplay, the children would also learn how mediRobbi could act or react, and how they can control the behavior of mediRobbi.

We designed a game, the *Stop-and-Go Food*, and a game map on a large, colorful, configurable map (see figure 6) for kids to use along with the robot. The game map creates a playground space for the young patient and his/her new friend, the mediRobbi. The robot serves as an encouraging interactive companion to facilitate play in the game space. This *Stop-and-Go* game introduces youngsters to a typical peanut allergy. While the robot travels along the predetermined path, it will stop each time at a decision point of *Stop* (peanut-related) food item. Each decision point has a rotating wheel that allows the child to select a given food item to present to the robot to resume motion. If the child chose the item that does not contain peanuts, the path would be connected and the robot can continue to move. If the choice is a food item that could cause allergy reaction, the path segment would be of red color and therefore the robot would not move. The goal of the game is to help the mediRobbi robot properly negotiate the food allergy constraints and find its way home.



Fig. 6. The game map was designed for the board game of the Stop-And-Go Game



Fig. 7. The pseudo children playing/waiting room inside the Child Behavior Lab

During this familiarization stage, the *Stop-and-Go* game requires the child to familiar with the robot by giving orders and to learn what feedback the robot would react. The interaction requires the motor skills of the robot pet to move along the path. The child learns that mediRobbi is capable of motion. The robot uses light sensor to

differentiate the black path from any motion-halting white spaces or red spots. The child learns that mediRobbi can be controlled to resume the motion. In short, the children can learn how to interact with the mediRobbi in this stage.

Third Phase – Stage of Real Examination. In the third phase, the participant started their pseudo examination process with our fake doctor in a pseudo hospital environment. We utilized the space in Child Behavior Lab where our young participants proceeded the first two phases in the playroom (see Figure 7) and the third phase in another divided space next to this playroom.

Instead of leading mediRobbi back home, the leading and following roles were reversed. The mediRobbi served as a guardian for the children and showed the young patient how long and how many steps of the examination process would be. The mediRobbi walked through a pre-designed channel of route visualizing the length of the whole visit. That the young participant can get a general idea of how long the exam would be. Also, their new friendly companion, mediRobbi, was there and accompanied them through the whole process.

5 Design Challenge and Discussion

One major challenge encountered during the implementation of Lego Mindstorms is that there's some limitation within the Lego device that we have to adjust our interaction design based on the robot behavior. The other major concern about the Lego Robot is its size and weight. Some of the young children seemed to have trouble holding the robot to bring the robot around. Besides, they really have to be gentle with the robot in order not to break it. This is also difficult for the young children. Future design should adjust the size and weight of the companion so that young children can carry the companion with them easily.

6 Summary

By moving through the three-phased process, from introduction phase to real examination, mediRobbi allows the pediatric patient to create a personalized playmate, to engage in a trust relationship with a comforting friend, and to learn about health issues through engaging play. These concepts were considered in the whole interaction design to help young pediatric patients overcome unknown anxiety and fear during their hospital visit.

References

- [1] Rasnake, L., Linscheid, T.: Anxiety Reduction in Children Receiving Medical Care: Developmental Considerations. *Journal of Developmental And Behavioral Pediatrics* 10(4), 169–175 (1989)
- [2] Farnum C.: How to Make Doctors' Visits Easier for Kids (March 24, 2010),
[http://www.suite101.com/content/
how-to-make-doctors-visits-easier-for-kids-a217717](http://www.suite101.com/content/how-to-make-doctors-visits-easier-for-kids-a217717)

- [3] How to Prepare Your Child for a Hospital Visit,
http://www.ehow.com/how_3939_prepare-child-hospital.html
- [4] Martin, F., Farnum, J.: Animal-Assisted Therapy for Children with Pervasive Developmental Disorders. *Western Journal of Nursing Research* 24(6), 657–670 (2002)
- [5] Levinson, B.M.: The dog as a "co-therapist",? *Mental Hygiene* 46, 59–65 (1962)
- [6] Pediatric Patients at Hays Med Get a Companion,
http://portal.haysmed.com/portal/page?_pageid=43,67257&_dad=portal&_schema=PORTAL
- [7] Lange, R. L.: I'll be O.K., First Printing. JoshCo, LLC (1997)
- [8] Coping Kits Comfort Children in the Emergency Department,
http://www.nursezone.com/nursing-news-events/more-features/Coping-Kits-Comfort-Children-in-the-Emergency-Department_20482.aspx
- [9] Levinson, B.M., Mallon, G.P.: Pet-Oriented Child Psychotherapy, 2nd edn. Charles C. Thomas Publisher (1997)
- [10] Robin, M., Bensel, R.T.: Pets and the Socialization of Children. *Marriage & Family Review* 8(3), 63 (1985)
- [11] Blue, G.F.: The Value of Pets in Children's Lives. *Childhood Education* 63(2), 84–90 (1986)
- [12] Bryant, B.K.: The Richness of the Child-Pet Relationship: A Consideration of Both Benefits and Costs of Pets to Children. *Anthrozoos: A Multidisciplinary Journal of The Interactions of People & Animals* 3(9), 253–261 (1990)
- [13] Lego.com Mindstorms,
<http://mindstorms.lego.com/en-us/Default.aspx>

Design of Shadows on the OHP Metaphor-Based Presentation Interface Which Visualizes a Presenter's Actions

Yuichi Murata¹, Kazutaka Kurihara², Toshio Mochizuki³,
Buntarou Shizuki¹, and Jiro Tanaka¹

¹ University of Tsukuba, 1-1-1 Tennodai, Tsukuba-shi, Ibaraki 305-8573 Japan

² National Institute of Advanced Industrial Science and Technology,

1-18-13 Sotokanda Chiyoda-ku Tokyo 101-0021 Japan

³ Senshu University, 2-1-1 Higashi-mita, Tama-ku, Kawasaki-shi,
Kanagawa 214-8580 Japan

murata@iplab.cs.tsukuba.ac.jp,

k-kurihara@aist.go.jp,

tmochi@mochi-lab.net,

{shizuki, jiro}@cs.tsukuba.ac.jp

Abstract. We describe the design of shadows of an overhead projector (OHP) metaphor-based presentation interface that visualizes a presenter's action. Our interface work with graphics tablet devices. It superimposes a pen-shaped shadow based on position, altitude and azimuth of a pen. A presenter can easily point the slide with the shadow. Moreover, an audience can observe the presenter's actions by the shadow. We performed two presentations using a prototype system and gather feedback from them. We decided on the design of the shadows on the basis of the feedback.

Keywords: presentation, graphics tablet devices, digital ink, shadow.

1 Introduction

Some presentation systems provide an ink annotation feature. A presenter can draw supplemental information or fix presentation slides on the spot using it. The presenter usually uses it with graphics tablet devices (GTDs), such as tablet PCs and pen displays.

However, earlier studies do not provide the following two features that we believe the presentation systems with GTDs should provide.

The presenter can easily point the screen. The presenter often points the presentation slides with his/her pointer or laser pointer. In presentation using GTDs, doing this is difficult because the presenter already have a pen device. This problem makes the presenter draw some attentional marks that includes deictic references to the slide elements [1]. Deictic references should be provided with pointing rather than

attentional marks because digital inks are persistent while the deictic references are impermanent. Hence presentations using GTDs should be designed for that the presenter can easily point the screen.

The audience can easily observe the presenter's action. The presenter's actions are an important way to convey his/her intention. Moreover, some of the presenter's actions help the audience to easily follow the presentation. For example, the audience can notice when the slide changes and when and where the annotations are drawn by watching the presenter try doing them. However, in the presentation using GTDs, the audience cannot observe these actions because the presenter operates on the input area of GTDs while the audience watches a large shared screen. Hence, presentations using GTDs should be designed to allow the audience can observe the presenter's actions.

Our goal is to achieve the aforementioned two features in presentations using GTDs. These two features help and improve communication. Thus, the presentation using GTDs should provide the two features.

Our contribution is to present the design of shadows that is suitable for pointing in the presentation using GTDs. Note that its design is based on feedback from actual presentations. The result of our study contributes to all the presentation systems using GTDs.

This paper consists of the following parts. First, there is a description of our approach. Next, we present a description of experiments with our prototype system that gathers feedback that is the basis for a description of the requirements of the design of shadows. There follows a description of the improvement that is based on the requirement. Finally, it describes related work and conclusion.

2 Approach

We took our cue for achieving the two features from overhead projectors (OHPs). In a presentation using the OHP, the presenter can easily point to the transparency by casting pen's shadow as well as drawing ink annotations. Moreover, the audience can observe these presenter's actions on the screen since these actions are also projected by the OHP.

On the basis of these observations, we designed an OHP metaphor based presentation interface and implemented it as Shadowgraph+. Shadowgraph+ visualizes the pen's current position and orientation as the pen-shaped shadow (Fig. 1). The presenter can point to the screen with the pen-shaped shadow and simultaneously draw annotations. Therefore, the audience can easily observe the presenter's action.

We developed Shadowgraph+ based on feedback gathered from the experiments on actual presentations. First, we developed experimental implementations and used them in presentations. We gathered feedback from the audience of the presentations and developed Shadowgraph+ based on them.

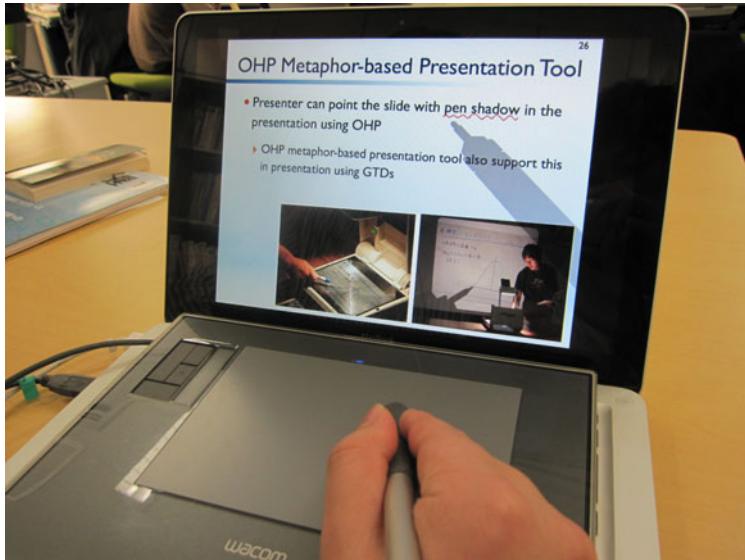


Fig. 1. Presentation using Shadowgraph+

3 Presentations Using the Experimental Implementations

We examined the experimental implementations in two presentations; one was at an academic conference and the other was as a university lecture. The implementations were add-ins of presentation software. We implemented two add-ins: one was for Microsoft PowerPoint and the other was for Kotodama [2] that is pen-based presentation tool. The add-ins superimposes a transformed ellipse as the pen-shaped shadow over the presentation slide. The procedure of the transformation is illustrated in Fig. 2. The add-ins transformed an ellipse using the pen's position, altitude φ and azimuth θ , all of which are the output of the GTD. The azimuth and the altitude of the pen's orientation are illustrated in Fig. 3. We used φ and θ for simulation of a real shadow. The shadow's color is $(R, G, B, A) = (0, 0, 0, 0.5)$. We used a GTD based on electro-magnetic sensing. The tablet can detect that the pen is held over the tablet even while it is not touching the tablet. This feature was needed to simulate casting the pen-shaped shadow.

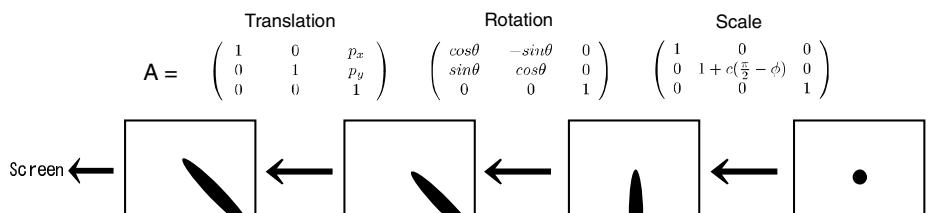


Fig. 2. Drawing procedure of the pen-shaped shadow

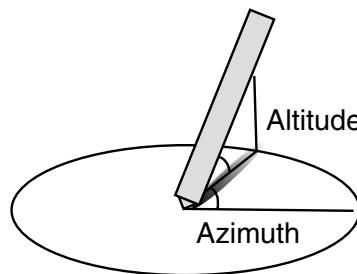


Fig. 3. Azimuth and altitude

3.1 Feedback from the Conference

We used our experimental implementation (PowerPoint add-in implementation) to make a presentation in a conference related on interactive systems and software. There were approximately 150 people in the audience. In the conference, almost all participants used a common chat system to hold discussions and to give the presenter feedback. We analyzed the logs of the chat system.

The number of comments was 148 in total. There were 91 comments mentioning our system or the presentation using it. The following were comments mentioned by various members of the audience.

- (A) Movement of the shadow is strange or distracts attention (10 comments).
- (B) Suggestion of compensation of wobbling (4 comments).
- (C) Suggestion of a feature that fixes the azimuth and the altitude (3 comments).
- (D) Shadow is pesky (2 comments).
- (E) Feel sick (2 comments).

The most interesting comments were related to the movement of the shadow. Comments (A) directly pointed out that the movement of the shadow is bad. Comments (B and C) implied it because they are suggestions of stabilizing the movement of the pen-shaped shadow.

The second point of interest in the comments implied that the shadow claimed too much attention. Comments (D) pointed this out directly. Comment (E) implied that the shadow claimed too much, and so the audience felt nauseous.

3.2 Feedback from University Lecture

One of the authors used the implementation for Kotodama in a lecture on teaching methodology of information sciences in the university. There were fourteen students. We introduced our system before the lecture and instructed the students to attend in usual.

After the lecture, we obtained feedback by questionnaire. The number of students who had positive comments and negative comments are shown in Table 1.

Table 1. Numbers of students who made positive comments or negative comments

		Number of Students who made	
		Positive comments	No positive comments
Number of Student who made	Negative comments	7	2
	No negative comments	4	1

The following were comments gathered from the questionnaire.

- (G) It is clear to see where the presenter is pointing (6 students).
- (H) It was clear to understand the point the presenter was talking about (3 students).
- (I) The shadow claims attention (5 students).
- (J) Appearance and disappearance of shadows claim attention (1 student).
- (K) Flickering shadows distract attention (2 students).
- (L) Eye is irritated (1 student)
- (M) Wiggled shadow distracts attention (1 student).

Eleven students in total made positive comments while nine students made a negative comment. Seven commented both of positively and negatively. One student made only a neutral comment. The comment only suggests another shape of the shadow, which was to use a detailed silhouette as a pen-shaped shadow.

The positive comments (G) indicate that the audience can well observe where the presenter is pointing. Comments (G and H) indicate that the shadow helps that the audience to better comprehend the course of the story. Because this comprehension is important for the presentation, this yields positive effects on the presentation. The negative comments (I and J) indicate that the shadow claimed too much attention. We found reasons from comments (J, K and M). These comments are related to the design of the shadow. We could ease these negative effects by improving the design of the shadow.

3.3 Design Guidelines Based on Feedback

From the feedback of the experiments, we found that the shadows support the presentation but too attract the audience's attention too strongly. More specifically, the unnatural orientation changes made the audience sick and distracted their attention. This result suggests that the shadow needs a subdued. Moreover, the negative comments gave us the following design guidelines:

Guideline 1 To reduce the visual strain, transparency should be higher unless the color is difficult to see.

Guideline 2 To reduce unnatural orientation changes of the shadow, they should be stabilized.

4 Implementation of Shadowgraph+

Following the design guidelines, we implemented Shadowgraph+. The previous design and the improved design of the shadow are shown in Fig. 4.

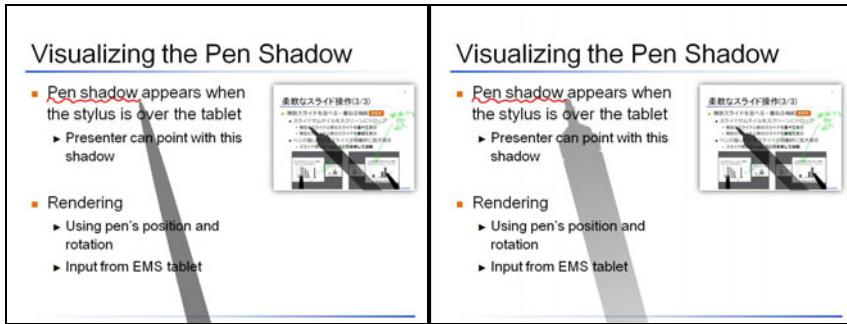


Fig. 4. Previous design (left) and improved design (right)

We changed the appearance of the shadow from a transformed ellipse into a gradational silhouette (Fig. 5). The color of the nib color is $(R, G, B, A) = (0, 0, 0, 0.4)$ and the tail color is $(R, G, B, A) = (0, 0, 0, 0)$. Following Guideline1, the gradation makes the place where the presenter is pointing obvious, but the shadows comparatively little attention.



Fig. 5. Gradational silhouette of pen-shaped shadow

Following Guideline 2, our Shadowgraph+ ignores the altitude and strongly filters the azimuth using a following low-pass filter:

$$\text{filtered } x_n = \frac{(\alpha - 1)x_{n-1} + x_n}{\alpha}$$

where x_n is the n-th azimuth value, and α is a parameter that represents the strength of the filter. We empirically determined the parameter as $\alpha = 50$.

5 Presentation Using Shadowgraph+

After refining the design of the shadows, one of the authors made a presentation. The audience of approximately 100 people included software developers, project managers, and researchers. We collected comments from the audience informally, and there are no negative comments about visual strain. This indicates our new redesign reduces the visual strain for the audience.

6 Discussion

From the aforementioned experiments, we got design guidelines and improved the design of shadows. While we have mainly discussed designing shadows and the result

is not enough to validate our approach. But some feedback reinforce that our approach worked as we intended. From the feedback (G), we conclude that the audience could observe where the presenter is pointing. This implies that the audience can observe the presenter's action.

7 Related Work

There are related works that support pointing by superimposed human bodies and tools, or that support computer-human interaction by superimposed shadows of human bodies and tools.

7.1 Support Pointing by Superimposed Human Bodies and Tools

The following research supports pointing by superimposing human bodies and tools on the computer workspace. Videodraw [3] and Clearboard [4] are distributed shared drawing system and superimpose the cooperator's body and his/her tools on the shared canvas. They support pointing and communication using it by the superimposing. C-Slate [5] supports the same feature on digital applications.

Videowhiteboard [6] and LIDS [7] superimpose shadows of the cooperators' body. While superimposing shadows have less textural information, it can convey the pointing and current situation of the cooperator in a calm.

Distributed Tabletops [8] and Video Arms [9] superimpose human arms extracted by detecting skin color. These systems also support pointing between cooperators in distant places. The detection can be implemented as software, and so the composition of systems can be smaller.

Meanwhile, our research supports the pointing in particular for presentations. Furthermore, it works with current presentation environments.

7.2 Support Computer-Human Interaction by Superimposed Shadows of Human Bodies and Tools

Wesugi et al. present an interaction between the tool's shadow and object in virtual space [10]. They describe that human perceive the tool's shadow that the human have as a part of the human's body, and they also describe the rationale for interaction by the tool's shadow.

Shomaker et al. present Shadow Reaching [11], interaction using the shadow of the human body. The shadow becomes bigger when the human is close to the shadow's light source. By making use of the natural nature of the shadow, Shadow Reaching achieves faster pointing of a distant object on a large screen.

Meanwhile, our research is mainly designed as pointing for the benefit of the audience rather than for interaction with computers.

8 Conclusion

We presented the design process of the shadow and implementation of the OHP-metaphor-based presentation interface with the improved shadow. We found that the

translucency of the pen-shaped shadow should be higher and the unnatural orientation changes should be stabilized.

References

1. Anderson, R., Hoyer, C., Wolfman, S., Anderson, R.: A study of digital ink in lecture presentation. In: The SIGCHI conference on Human factors in Computing Systems, pp. 567–574. ACM, New York (2004)
2. Kurihara, K., Igarashi, T., Ito, K.: A Pen-based Presentation Tool with a Unified Interface for Preparing and Presenting and Its Application to Education Field. Computer Software 23(4), 14–25. JSSST (2006)
3. Tang, J.C., Minneman, S.L.: Videodraw: A video interface for collaborative drawing. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 313–320. ACM, New York (1990)
4. Ishii, H., Kobayashi, M.: Clearboard: A seamless medium for shared drawing and conversation with eye contact. In: Proceedings of the SIGCHI Conference on Human factors in Computing Systems, pp. 525–532. ACM, New York (1992)
5. Izadi, S., Agarwal, A., Criminisi, A., Winn, J., Blake, A., Fitzgibbon, A.: C-slate: A multi-touch and object recognition system for remote collaboration using horizontal surfaces. In: International Workshop on Horizontal Interactive Human-Computer Systems, pp. 3–10. IEEE Computer Society, Los Alamitos (2007)
6. Tang, J.C., Minneman, S.: Videowhiteboard: video shadows to support remote collaboration. In: Proceedings of the SIGCHI Conference on Human factors in Computing Systems, pp. 315–322. ACM, New York (1991)
7. Apperley, M., McLeod, L., Masoodian, M., Paine, L., Phillips, M., Rogers, B., Thomson, K.: Use of video shadow for small group interaction awareness on a large interactive display surface. In: Proceedings of the Fourth Australasian User Interface Conference on User Interfaces, pp. 81–90. Australian Computer Society (2003)
8. Tuddenham, P., Robinson, P.: Distributed Tabletops: Supporting Remote and Mixed-Presence Tabletop Collaboration. In: International Workshop on Horizontal Interactive Human-Computer Systems, pp. 19–26. IEEE Computer Society, Los Alamitos (2007)
9. Tang, A., Neustaedter, C., Greenberg, S.: Videoarms: Embodiments for mixed presence groupware. In: People and Computers XX —Engage, pp. 85–102. Springer, Heidelberg (2007)
10. Wesugi, S., Kubo, T., Miwa, Y.: Tool-type interface system supporting for an expansion of body image toward a remote place - development of virtual shadow interface system. In: SICE 2004 Annual Conference, pp. 912–917. Society of Instrument and Control Engineers (2004)
11. Shoemaker, G., Tang, A., Booth, K.S.: Shadow reaching: a new perspective on interaction for large displays. In: Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology, pp. 53–56. ACM, New York (2007)

Web-Based Nonverbal Communication Interface Using 3DAgents with Natural Gestures

Toshiya Naka^{1,2} and Toru Ishida²

¹ Panasonic Corporation, 3-1-1 Yagumo-naka, Moriguchi City, 570-8501 Japan

² Graduate School of Informatics Kyoto University, Yoshida Honmachi
Sakyoku, Kyoto, 606-8501 Japan

naka.tosiya@jp.panasonic.com, ishida@i.kyoto-u.ac.jp

Abstract. In this paper, we assumed that the nonverbal communication by using 3DAgents with natural gestures had various advantages compared with only the traditional voice and video communication, and we developed the IMoTS (Interactive Motion Tracking System) to verify this hypothesis. The features of this system are that the natural gestures of 3DAgents are captured easily by using interactive GUI from the 2D video images in which some characteristic human behaviors are captured, transmitted, and reproduced by natural gestures of 3DAgents. From the experimental results, we showed that the accuracy of captured gestures which often used in web communications was within the level of detectable limit. And we found that human behaviors could be characterized by the mathematical formula, and some of the information could be transmitted, especially some personalities such as quirks and likeness had the predominant effects of impressions and memories of human.

Keywords: Computer graphics, Agent, Virtual reality and Nonverbal Communication.

1 Introduction

According to the rapidly development of web-based ICT technologies, web users exceeded 30 million, about 25 percent of the world population over the past decade. During this time, many new communication services and social activities in web applications have actively proposed. In the near future, web communication will be expected to accelerate more and more in diversification and borderless. On the other hand, it began to become apparently the specific problems such as cultural differences and digital divide. In addition to the traditional voice and/or video based communications, the effective use of 3DCG and multimedia will become important for the intercultural communication and collaboration [1]. Not only limited on the web but also the communication in real world, it has been previously pointed out that the importance of nonverbal information such as facial expressions, gestures and behavior which were expressed in addition to verbal communication, Mehrabian has studied that about 55 percent was dependent by nonverbal information in communication, and the transmission of information through language was less than seven percent [2]. In the web communication, there are critical aspects of anonymity

such as exchanging information with strangers, cooperative works and shopping, which are the special environment as virtual space. Under these situations, we examined the effectiveness of web-based 3DAgents communication with natural gestures.

2 Background

When we are trying to utilize new multimedia for the web communications, we assumed that it would become very important of 3DAgents communication with natural gestures equivalent to the real world. According to the web-based 3D, web3D technologies begun to study since the late 90s, and they have standardized for the basic part as VRML2.0 in 1997 [3]. After that, it was extended with multi-agent function and Humanoid animation [4]. According to the previous research for expressing the natural 3DAgents in the field of computer graphics, There are many studies such as Witkin [5] considered the operation of physical constraints to express the dynamics and Badler [6] proposed new method for accurate representation of the complex behavior using dynamics and Kudoh [7] proposed some methods for generating appropriate behavior against the balance disturbance. In order to express the natural behavior of the 3DAgents, motion capture is one of the direct ways to get humans actions, but this method is usually required the accuracy of calibration and the special environment to capture the motions of our everyday life. On the other hand, there are many researches in which how to capture the humans actions from 2D video using image processing techniques such following research as, Hoshino [8] studied for match moving technique that estimated 3D motions from the movement of the feature points on 2D images, and Wei [9] also estimated the motions by using contact constraints and Newtonian physics from the feature points of monocular video. But in many previous researches, there were a few studies in which treated from the simple motion tracking method to real-time transmission, and referring to the effectiveness when using 3DAgents for the web communications. In this paper, we developed the end to end platform in which we could capture some motions of human, transmit them in real-time via the web and reproduce those natural motions of 3DAgents. By using this system, we examined the effectiveness of 3DAgents communication which focused on the following items.

Characteristic gestures of human can be transmitted. As shown in Figure 1, when we faithfully mimic some human gestures and if we use 3DAgents for the non-verbal communication, whether we can translate the information of beyond words "human personality (e.g. emotion and like/dislike)" or not.

Characteristic gestures of human can be symbolized and stored in our memory. Some characteristic gestures of human whom we like and respect to her (such as Mentor) are clearly affecting the information transmission and memory of human. To verify the above mentioned hypotheses, we constructed 3D motion tracking system named IMoTS, in which some natural motions can be estimated by 2D image tracking using interactive GUI, transmitted in real time via the web and realized as the real actions of 3DAgents. We summarize the feature of IMoTS below.

1. 3DAgent is defined by the hierarchical structured nodes, and his natural postures and behavior are controlled by the variable accuracy using kinematics.

2. Using 2D video images in which some characteristic human gestures are captured; user can easily plot some virtual markers on each joints position using interactive GUI. Then we estimate 3D motion from optical flow of those virtual markers using time-space gradient method. In order to minimize both estimation error and affection of missing data by occlusion, we can compensate for animation data using IK.

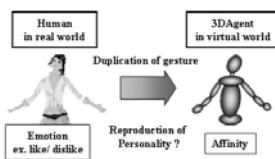


Fig. 1. Nonverbal communication with 3DAgents

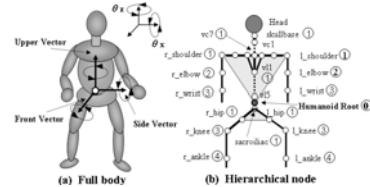


Fig. 2. 3D Skeletal structure and nodes

3 Nonverbal Communication Platform IMoTS

In this Section, we describe the basic principles and overall structure of nonverbal communication platform IMoTS. Using this system, we can capture the natural gestures of human, transmit them in real-time on the web and generate 3DAgent with natural gestures. IMoTS is constructed by two main functions, first one is the function of tracking and estimating some 3D motions from 2D video images, and second one is the operation of real-time transmission of 3D motion data streaming [10]. We describe about only the former feature in this Section.

3.1 Hierarchical Skeletal Structure and Motion Control of 3DAgent

In IMoTS system, we define the hierarchical skeletal structure of 3DAgents shown in Figure 2. They can be defined similarly in both joints of the fingers and toes [4]. In figure 2 (b), the body triangle which including the vertices of Humanoid root and both shoulders, and the hip triangle which including with vertices of Humanoid root and both hips represents the initial state of 3DAgents posture. Using this hierarchical structure, we control the natural posture and motions of 3DAgents using kinematics.

We calculate the posture from the direction of the higher in the hierarchy node from such as $r_shoulder$, r_elbow to r_wrist using direct kinematics shown by equation 1. In the equation 1, q_{i+1} is the direction vector of $i+1$ th lower joints, and q_i represents the direction of the higher joint i th and N describes the number of degrees of freedom for all joints.

$$q_{i+1} = f(q_i) \quad \text{for } i = 1 \sim N-1 \quad (1)$$

3.2 Interactive Motion Tracking

Next, we explain how to capture the natural motions of 3DAgent in more detail. In order to calculate some pose of 3DAgent, we have to set q_i of equation 1 as some of

time series data for each joints angle which is consisted by the rotation angle ($\theta_x, \theta_y, \theta_z$) in XYZ axis. On the other hand, for the central position node as Humanoid root of 3DAgent, we must set both the orientation and rotation vector in world coordinates.

3.2.1 Sequence of 3D Motion Data Tracking

IMoTS can reproduce the natural gestures of 3DAgent by setting q_i to some characteristic behavior of the 3D data which is captured from humans gestures. Typically, we can be used to accurately obtain the behavior of human using motion capturing. But existing motion capture systems are usually expensive, difficult to calibrate and have some problems such that limited to indoor environments and settings and so on. On the other hand, IMoTS is able to easily capture the 3D motions of variety gestures which are taken from 2D camera images [8], [9]. Figure 3 shows the basic sequence of motion capturing operation of IMoTS.

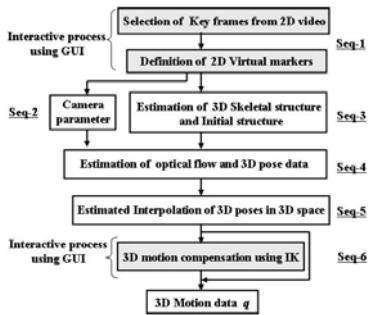


Fig. 3. Sequence of 3D motion tracking

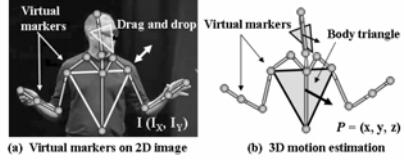


Fig. 4. Estimation of 3D motion from 2D video images (a) shows virtual markers which selected on the joints position of human body by using interactive GUI

Seq-1: Selection of key frames and settings virtual markers by using interactive GUI: In the first step, to estimate 3D motions we need to select some key frames in which some characteristic gestures of humans are taken by 2D images. Although the many number of selecting key frame must increase the accuracy of operation, but they are trade-off between the efforts of GUI manipulations. We will verify the details of this relationship by the experiment in Section 4. In the next step, we have to set the virtual markers to the corresponding joints position of each skeletal structure in Figure 2 (b), for those selected key frames images. In this operation, user can set or modify the position data of those virtual markers for each key frame by using GUI operations.

Seq-2: Estimation of the camera parameters: Then we estimate the camera parameters ρ from 2D images by using the technique of active appearance models [12] which is often used in the fields of computer vision.

Seq-3: Estimation of initial skeletal hierarchy of 3DAgents: In order to estimate the initial location of the skeletal structure as shown in Figure 2 (b), IMoTS estimates the displacement of the body triangle including vertices of Humanoid root and both shoulders, and/or the hip triangle including ones of Humanoid root and both hips. We can estimate the initial value of them from the displacement between the direction vector of the body triangle (counterclockwise direction is positive) and z-direction

(perpendicular to the image shown as Figure 2 (b)), and estimate from the tilt angle of each side of those 2D triangles from its initial position. In this estimation, we have to use the link-length ratio of the average adult's value, as for the links length of the skeletal structure shown in Figure 2 (b).

Seq-4: Estimation of 3D motion data using Spatio-temporal gradient method: In the next step, we estimate 3D motion data q using the optical flow of virtual markers at each joint in the images of selected key frames. When we define I as the brightness values of any of virtual markers which coordinates are defined by (X, Y) as shown in Figure 4 (a), the relationship with the corresponding point of the 3D skeleton model $p = (x, y, z)^T$ can be represented by the following formula in equation 2, using the gradient-based optimization method.

$$\begin{aligned} K(\partial x / \partial t, \partial y / \partial t, \partial z / \partial t)^T + \partial I / \partial t &= 0 \\ K = \{f(Ix/z), f(Iy/z), (xIx + yIy)/z\} \end{aligned} \quad (2)$$

In addition, if we define the rotation vector of each skeletal joint of 3DAgents and the position vector of Humanoid root joint as q , then the relationship of equation 2 can be expressed by equation 3 by using the Jacobian $J(q)$. Then we are able to estimate the 3D motion data by solving q of equation 3 [8].

$$K J(q) (\partial q / \partial x) + \partial I / \partial t = 0 \quad (3)$$

Seq-5: Estimated interpolation of motions in 3D space: We have to compensate the 3D motion data in between key frames by interpolating operation using 3D motion data q values of each joint which are mentioned in Seq-4. In this process, we effectively use the predicted velocity data in 3D angles which are estimated by using optical flow of each joints angle which mentioned in Seq-4.

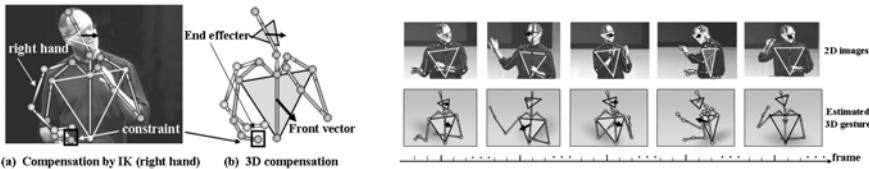


Fig. 5. Compensation GUI operation of 3D poses by using IK

Fig. 6. An example GUI of motion tracking using IMoTS

Seq-6: Compensation of tracking motion using inverse kinematics (IK): When we will estimate 3D motions from the commonly used 2D video images, these values usually contain some errors which are occurred by occlusion, human errors of determining the virtual markers position. For this reason, we use IK to correct these errors. Assuming the case such as “feet sliding on the floor” or “hand with an object breaks into/not contact with an object”. These errors are usually significant and partly can not be ignored, so we have to correct the position by using the constraint condition of the end effector as equation 4. In this equation, J^+ is named inverse Jacobian and it is obtained by minimizing the sum of squares as $\Delta q \Delta q^T$. According to the above

mentioned constraints, user can manipulate the position of the end effectors by drag and drop operation using the interactive GUI in 3D space. In Figure 5, we illustrate an example GUI operation.

$$\Delta q = J + \Delta p : \quad J^+ = J^+ (J J^+)^{-1} \quad (4)$$

4 Experiments and Discussion

In this section, we evaluated the basic performance of 3DAgent-based nonverbal communication platform IMoTS and discuss some basic experimental results by using IMoTS to examine the effectiveness of nonverbal communication with 3DAgents with natural gestures.

4.1 Evaluation of 3D Motion Tracking by IMoTS

Firstly, we examined the accuracy of 3D motion data tracking operation of IMoTS. Generally speaking, the key factors affecting the accuracy of IMoTS are considered such as (a) human errors in the GUI operation when user setting up the virtual markers in Seq-1, (b) 2D errors which will occur depending on the physical conditions such as lighting and camera position when taking pictures and (c) some errors due to the operation of estimating algorithm described in Seq-2 through Seq-5.

Among these key factors, it will be estimated that the interactive process of users operations in Seq-1 will significantly affect on the estimation accuracy of IMoTS. We examined to evaluate quantitatively the relationship between estimated accuracy which depending the selection of the type of gestures and numbers of key frame.

Study-1: Figure 6 shows an example of the estimated 3D poses using IMoTS of each selected key frame. The upper images in Figure 6 show some scenes of setting the virtual markers on each key frame images by using GUI operation (See Seq-1 in section 3.2). And the lower ones show some 3D poses estimated corresponding posture of 3DAgent using IMoTS algorithms described in 3.2. Including another all gestures which used in our experiments, we confirmed that the accuracy of estimation was within the level in which errors could not be detected visually.

4.1.1 Accuracy of Relation between Type of Gesture and Physical Condition

As the next experiment, we verified the accuracy of animation data which was interpolated in between the selected key frames using both estimated velocity q and prediction techniques which mentioned in Seq-4. Not only IMoTS, but also for other tracking systems, the accuracy of normal tracking operation is largely determined by the type of gestures (especially the speed of motion at each joint) and the shooting conditions. In the following experiments, we classified into three main types of gesture groups of human's motion, and we examined the relationship between the accuracy of motion tracking and some of the representative behavior of each group.

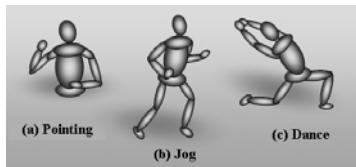


Fig. 7. Example of representative gestures of three groups

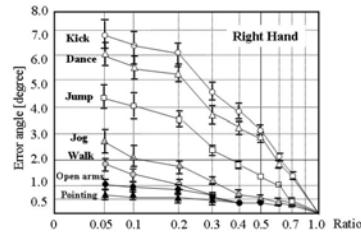


Fig. 8. Relationship between the ratios of selected key frame and IMoTS error (right hand joint)

Types of gesture : Figure 7 shows some example of the representative human behavior of each group which used in our experiments.

(a) *Group of gesture using upper body action:*

Many gestures of this group are relatively small amount of displacement of the body triangle, and these gestures are usually represented by the movement of links of upper arms, body and/or neck. This motion group is often used in the web communication such as speech and interactive conversations. We selected five representative gestures in the following experiments such as "spread both hands (open arms)", "raise one hand" and "pointed by one hand (pointing)" .

(b) *Group of gesture using full body action and basic action in usual communication:* Many of the gestures of this group are realized by the movement of Humanoid root joint in addition to the actions of group (a). This type of action usually accompanies the movement of the body such as the situation of moving on the stage with the body action. Lighting conditions are relatively severe and shooting conditions are zoomed out. In case of the action of this group, displacements of the body triangle and/or the hip triangle are often much larger than the group (a). We used five basic gestures such as "walk", "jogging" and "thinking by walk" in the following experiments

(c) *Group of gesture with full body actions:*

These gestures are represented using all the links of body such as dancing and playing sports, and movements of each links are usually large. Both conditions of lighting and resolution are usually severe which compared with gesture group (a) and (b). We used five representative actions such as "kick", "jump" and "dance" in our experiments.

Condition of lighting and camera position: For the gestures of the above mentioned three groups, we illuminated the actor by lighting from two direction of top of left and right front position. In addition, we captured each action from the camera position of front direction from her. Moreover the effective resolution of 2D image is 640×480 pixels and we set frame rate as 30 frames/sec. We compared the RMS error value at each joints from the above mentioned reference values as Δq which was the difference in 3D space.

Study-2: Figure 8 shows a relationship between the number ratios which selected key frames and RMS error of estimated 3D motion of above mentioned three types of gestures. In Figure 8, we plotted some experimental results of seven representative gestures of group (a) "open arm" and "pointing", group (b) "jogging", "walk" and group (c) "kick", "dance" and "jump" individually. The horizontal axis is the ratio of

the number of selected key frames, and vertical axis shows RMS error value of each joint angle. Figure 8 shows “right hand” values which are the maximum RMS error (the worst value of all joints) within all links of 3DAgent. From these experimental results, we found the following trends. The smallest RMS error group of estimation was the upper body gesture group (a), and tended to increase RMS errors from group (b) to (c). We estimated that this result was caused by selecting the velocity values during interpolation operation for the prediction of each transfer. So the accuracy of prediction was relatively higher such gestures as the velocity of each joint’s angle changed monotonically in 3D space. On the other hand, the prediction errors increased greatly for such gestures as difficult to predict in interpolation between key frames (e.g. kick, jump and dance). For almost all gestures used in our experiments, when we select the ratio of number of key frame more than 0.4 (equals to sampling more than 4 frames from 10), then the maximum RMS error of IMoTS could be reduced from 3.5 to 4.0 degrees. And for the group gestures (a) which are commonly used in the normal interactive web communications, even if the ratio of key frame is approximately 0.1, the maximum cumulative RMS error is kept within from 1.5 to 1.0 degree. These values are practically enough to use for our evaluation purpose.

4.2 Experiment of the Effectiveness with 3DAgents Communication

We conducted some basic experiments to verify the effectiveness of 3DAgents communication. Firstly, we considered to confirm that distinctive gestures of human can be transmitted and some of the gestures can be easily stored in our memory.

Question-1: Whether some of the distinctive human gesture can be translated.

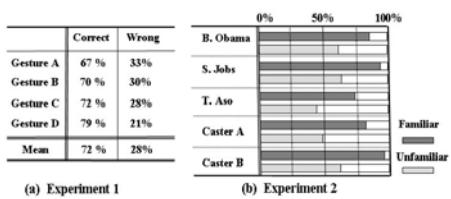


Fig. 9. Experimental result of 3DAgents communication

selected total of 20 subjects of adult men and women aged 20 to 50 years. As shown in Figure 10, in advance of evaluating the gestures of 3DAgents, we showed 2D images as twice in succession (approximately 30 seconds length and 10 types) in which human’s actions captured from some distinctive speech and dialogue scenes. Then after a few 10 seconds, we showed each subject to some reproduced 3DAgents gestures which were randomly selected among the actions of those 10 types. In these experiments, 3DAgent had no verbal information and we asked each subject to compare the gestures with only nonverbal information which captured faithfully. Each subject selects one of the original 2D images (A to D) from 10 types of images. We evaluated whether each distinctive humans gesture such as A to D could be translated by using the percentage of correct answers (that is matching to the corresponding gesture images).

We did some basic experiments to verify that some of the human personality (e.g. habits, likeness and feeling) were able to pass by 3DAgents who mimic the gestures of human in faithfully as shown in Figure 1. We had the hypothesis that nonverbal communication using 3DAgent could translate some of the personalities which were difficult to be passed by only verbal communication.

In the following experiments, we

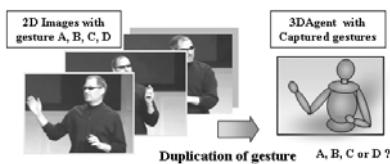


Fig. 10. Evaluation of 3DAgents communication

Study-3: Figure 9 (a) shows an example of the experimental results. We chose the distinctive typical human gestures A, B, C and D in 30 seconds such as "shaking hand", "pointing", "open arms", and "raise hand" individually. The percentage of each correct answer rate was 67 to 79 percent. From these results, we can say that some distinctive gestures tend to have the advantage of our memory or impression.

Question-2: Whether some distinctive gestures of human can be quantified and easily stored in memory.

Next, we did some basic experiments to verify some characteristic gestures of the human who are respected by some subjects (such as Mentor) had some clear affection to information transmission. We conducted some subjective test under the same conditions as *Question-1*. In this experiment, we selected the typical gestures of both types (X) who you like or familiar persons with the subjects, and (Y) strangers or not-familiar ones. Figure 10 shows an example of representative familiar persons such as (a) Prime Minister of Japan (T. Aso) and (b) President of U.S. (B. Obama). In addition, we chose celebrities such as S. Jobs, two Japanese and American TV anchors indicated by Caster-A and Caster-B. Using same subjects in the experiment of *Question-1*, we showed these selected gestures to 7 or 8 subjects who were "familiar", and also showed them to 5 or 6 subjects who were "not interested".

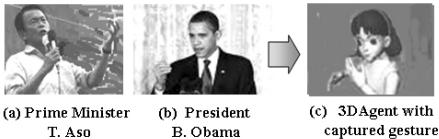


Fig. 11. Distinctive gestures of familiar persons (a), (b) and reproduced gesture by 3DAgent (c)

In the experiments, each subject was shown about 30 seconds video that taken those gesture and we showed them two times successively. After 10 seconds from showing, we evaluated by the ratio of correct answer which seemed to be trying to choice the same reproduced gesture of 3DAgent. Figure 9 (b) shows an example of the

experimental results. In Figure 9 (b), we expressed the ratio of correct answers that are "familiar" by black line, also expressed "non-familiar" by gray respectively.

Study-4: As the results shown in Figure 9 (b), the percentage of correct answers of the subjects group who are familiar with exceeded 75 percent and the correct answers were 90 percent or more of celebrities such as S. Jobs and especially anchor Caster-B. On the other hand, the percentage of correct answers of the subjects group who were non-familiar with settled within 50 to 65 percent. From this experiment, we conclude that some gestures whom people like or familiar had some advantages for their memory or impression. Moreover, the result was lower percentage of correct answers than the other group of non-familiar such as T.Aso and anchor Caster-B. We estimated that the gestures of Japanese were usually smaller than those of Westerners, and such difference was caused by the difficulty of distinguishing with gestures of others even if the same behavior.

5 Conclusion

In this paper, we developed the IMoTS to verify the effectiveness of web-based 3DAgents communication with natural gestures. The features of this system are that the natural gestures of 3DAgents are captured easily by using interactive GUI from the 2D video images, transmitted over the web in real-time, and reproduced by the natural gesture of 3DAgents. By using IMoTS system, we found the following conclusions. According to the tracking accuracy of 3D gestures of IMoTS, the estimated maximum RMS errors of 3DAgents were less than the detective limit of key frame selection rate between 0.05 and 0.1. IMoTS system had the sufficient accuracy in the practical use of our evaluation purpose. In the nonverbal communication using 3DAgents with natural gestures, some of the characteristic gestures could be expressed by the mathematical formula, and transmitted some of the personality such as quirks and likeness. When we used some distinctive gestures which we were “familiar” or “respect”, there was clear superiority in memory or impression which compared to the gestures of “non-familiar”. In the future, we would like to improve the accuracy of IMoTS system, and we hope to extract key factor in nonverbal communication when we use 3DAgent of general communication purpose.

References

1. Reeves, B., Nass, C.: *The Media Equation*. Cambridge University Press, Cambridge (1996)
2. Mehrabian, A.: *Silent messages: Implicit communication of emotions and attitudes*, 2nd edn., Wadsworth, Belmont, California (1981)
3. The Virtual Reality Modeling Language, ISO/IEC DIS 147721 (1997)
4. Humanoid animation (H-Anim), ISO/IEC 19775-1 (2008)
5. Witkin, A., Kass, M.: Space-time constraints. In: Proc. 15th Annual Conference on Computer Graphics and Interactive Techniques, pp. 159–168 (1988)
6. Badler, N., Phillips, C., Webber, B.: *Simulating humans: computer graphics, animation and control*. Oxford University Press, Oxford (1993)
7. Kudoh, S., Komura, T., Ikeuchi, K.: The dynamic postural adjustment with the quadratic programming method. In: Proc. 2002 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 2563–2568 (2002)
8. Hoshino, J., Saito, H., Yamamoto, M.: A Match Moving Technique for Merging CG Cloth and Human Video. *Journal of Visualization and Computer Animation* 12(1), 23–29 (2001)
9. Wei, X.K., Chai, J.-x.: Interactive tracking of 2D generic objects with spacetime optimization. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) *ECCV 2008, Part I. LNCS*, vol. 5302, pp. 657–670. Springer, Heidelberg (2008)
10. Naka, T., Mochizuki, Y.: High speed motion data transmitting technique for humanoid animation on web-based cyberspace. *IEICE J94-D(5)*, 1–9 (2011)
11. Wei, X., Chai, J.: VideoMocap: Modeling Physically Realistic Human Motion from Monocular Video Sequences. *ACM Transactions on Graphics* 29(4) (2010)
12. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. In: Burkhardt, H., Neumann, B. (eds.) *ECCV 1998. LNCS*, vol. 1407, p. 484. Springer, Heidelberg (1998)

Taking Turns in Flying with a Virtual Wingman

Pim Nauts, Willem van Doesburg, Emiel Krahmer, and Anita Cremers

TiCC, Tilburg Centre for Cognition & Communication

Tilburg University, The Netherlands

P.O. Box 90153, 5000 LE Tilburg

{P.A.H.Nauts,E.J.Krahmer}@uvt.nl

PCS, Perceptual and Cognitive Systems,

TNO, The Netherlands

P.O. Box 23, 3769 ZG Soesterberg

{willem.vandoesburg,anita.cremers}@tno.nl

Abstract. In this study we investigate miscommunications in interactions between human pilots and a virtual wingman, represented by our virtual agent Ashley. We made an inventory of the type of problems that occur in such interactions using recordings of Ashley in flight briefings with pilots and designed a perception experiment to find evidence of human pilots providing cues on the occurrence of miscommunications. In this experiment, stimuli taken from the recordings are rated by naive participants on successfullness. Results show the largest part of miscommunications concern floor management. Participants are able to correctly assess the success of interactions, thus indicating cues for such judgment are present, though successful interactions are better recognized. Moreover, we see stimulus modality (audio, visual or combined) does not influence the ability of participants to judge the success of the interactions. From these results, we present recommendations for further developing virtual wingmen.

Keywords: Human-machine interaction, turn-taking, floor management, training, simulation, embodied conversational agents, virtual humans.

1 Introduction

In military aviation there is a tendency towards replacing manned aircraft by Unmanned Aerial Vehicles (UAVs) in hybrid teams (a human lead pilot assisted by a UAV). In the future, human wingmen¹ are foreseen to be replaced by cognitive systems (virtual wingmen) that function as autonomous members in flight operations. Since fighter pilots are highly-skilled, highly-trained professionals, replacing a human wingman by a cognitive system puts high demand on the system's ability to resemble the replaced human in both appearance and behavior.

Given that one of the most important aspects of flying missions in teams is reciprocal trust (the basis of which is a clear understanding among team members) the specific challenge of developing a virtual wingman is to provide the wingman with an

¹ A wingman is the subordinate of the lead in a two-man (*two-ship*) formation.

understanding of how to interact with its human team members, how to avoid potential misunderstandings and how to solve them when they occur. We argue that, to achieve this, it is important to first look closely at how humans detect and repair miscommunications in this particular setting. This study adds to existing literature in problem detection and -repair by investigating the interaction between a highly skilled operator, a pilot, and a virtual interlocutor, his virtual wingman.

1.1 Embodied Cognition

As a pilot, you want to be assured that whatever happens your partner, either your wingman or your lead, will be up to the task at hand without any need for directing extra attention to a clear understanding among partners. Replacing a human wingman by an UAV, presented as a virtual wingman, thus asks for a cognitive system that can autonomously interact with the lead and other partners in flying as effectively as possible. Establishing the trust and confidence needed in aviation requires a virtual wingman to provide the suggestion of human-like communication capabilities [1] and to meet expectations of a pilot with regard to its specific role (e.g. domain knowledge, reliability).

The specific challenge then becomes to 1) understand the expectations raised by the suggestion of human-like communication and the suggestion of a wingman, 2) to understand what miscommunications occur and if they are related to either source of expectation, 3) to design dialogue repair strategies that stimulate acceptance and trust of the embodied cognition of the UAV. It is thus an effort of joining advances in computer science, specifically artificial intelligence, and insights from social sciences, capturing the richness and dynamics of human behavior in cognitive systems [2].

The level of humanness that human-machine interfaces show has increased over the years, from simple and static agents to virtual characters that interface between and bridge virtual and real worlds, e.g. for training and simulation in military contexts [3] [4] [5]. Such ‘virtual humans’ are defined as cognitive systems that look like, act like, and interact with humans but exist in virtual environments [2]. In other words, they are a virtual extension of a real-world entity (e.g. an UAV). Despite reported experiences with virtual humans there is no consensus on how to build an embodied cognitive system that can live up to the promise of human-like conversation.

1.2 Taking Turns

Natural language dialogue is a very important aspect of humanness a virtual wingman needs to master, as conversation is so defining of human interaction. However, human dialogue is not without errors and problems [6]. Such problems are best viewed as miscommunications - whatever the communicator was trying to convey is not understood by the addressee. Assuming miscommunications negatively affect partners’ satisfaction and the perceived effectiveness and trust and control so important to military aviation, these errors should be recognized and handled by a virtual wingman. They should resemble human repair mechanisms, specifically in a way that yields the highest satisfaction to the human partner [7] [8]. Such repair mechanisms, the ability to indicate when communication goes awry and react to these indicators, are an important aspect of our conversational abilities [10]. They are rich

by nature and use multiple cues to signify the occurrence of miscommunications (e.g. linguistic features, facial display, timing) [6] [9]. This information is derived from cues in nonverbal channels to control the flow of conversation [11]. With conversational partners posing and answering questions or statements, a sequence of turn-taking is present in the conversation and controlled using these cues, e.g. an end-of-utterance is pre-signaled by the speaker, indicating a current turn is about to be finished [12]. Giving up a turn is often signaled by looking at the addressee [2].

As turn-taking is part of how humans engage in conversation, it is an important characteristic to mimic in a virtual agent - it determines how natural a conversation will be perceived by the human partner. However, it is not clear whether the rules involved in the floor management of casual conversation extend to our specific domain. As such, in order to let virtual wingmen respond appropriately we first need to provide them with an understanding of how to manage the floor or detect error conditions [13] and when to direct extra effort towards human partners' understanding.

In search for a method to provide this understanding we can best look at how humans, specifically pilots, tackle these problems - supposedly by recognizing and processing cues the conversational partners elicit. As shown in other studies (e.g. [12] [14] [15]), perception experiments can reveal whether or not it is evident these cues can be derived from the conversation. We thus investigate if information on the flow of conversation is present in the interactions between pilot and virtual wingman.

To this end we set up a perception experiment based on perceptual judgment by participants. The experiment is preceded by an inventory of type of interactions (both miscommunications and successful interactions) present in interactions between a virtual wingman and a human pilot.

2 Data Collection

The data used in the analysis and experiment are taken from an earlier experiment within a pilot performance program, conducted to evaluate the attitude of military pilots towards working with a virtual wingman. The character used in the experiment, Ashley, is a female virtual agent under development at TNO Perceptual and Cognitive Systems (The Netherlands). An exhaustive description of Ashley's design is beyond the scope of this paper, important in that respect is Ashley is a life-like virtual agent in that she can mimic a genuine human in her appearance and behavior and can engage in dialogue on a level sufficient to test the context in interaction with human pilots. In the previous experiment, fighter pilots from the Dutch Air Force were assigned to fly a mission together with Ashley from front to back, i.e. practicing with Ashley on executing a *lost wingman procedure* with a pre-flight briefing, a (simulated) actual flight and post-flight debriefing. Their task was specifically in the role of the lead pilot with Ashley being their young wingman in training certain basic flight maneuvers and safety procedures.

3 Data Analysis

Data from the aforementioned experiment were first analyzed for occurrences of miscommunications within the conversation, dividing the raw material into separate

interactions (a set of turns that belong to answering a single question, including backchannel behavior) and evaluating each interaction for effectiveness. All selected videos shared the viewpoint (pilot visible, side-frontal; Ashley out of frame) and concern the debrief. In over two hours of video (02:03:01) with seven different pilots a total of 394 interactions were observed.



Fig. 1. Left: Example frame from the analyzed videos. People in the background do not engage in the interaction between Ashley and the pilots. Right: The set-up of the briefings. The pilot faces Ashley on the computer screen. A microphone detects speech, speakers are integrated in the room.

3.1 Coding

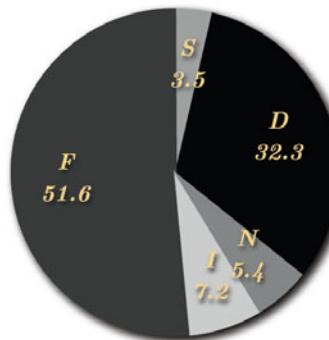
Each interaction present in the selected recordings was coded and scored from the perspective of the human partner (pilot) as we aim to derive cues from the pilots that might indicate an interaction is failing. Every set of turns, in which a question is posed by one partner and answered by the other, is defined an *interaction*. As strictly coding every set of turns disregards backchannel behavior, backchannel-cues from both partners are included. In most cases, a clear categorization was evident; when in doubt interactions were excluded. See below for an explanation of the coding. Only ‘first level’ problems are coded, the first occurring problem in an interaction, e.g. if a delayed answer *D* is followed by a nonsensical answer *S*, only the delay is coded. ‘Unnatural’ in a *Delay* refers to the perception of the coder. As a strict threshold for length (e.g. 3 seconds) disregards context it does not make sense to use an arbitrary measure (a request for simple confirmation requires less time to respond to than does an answer that requires cognitive effort).

3.2 Results

A first indicator of Ashley’s performance is the number of positive cases identified. From 394 identified interactions roughly half is successful (*F*, 51.6%). One third of all cases concern delayed responses (*D*, 32.3%), non-responses account for five percent (*N*, 5.4%), interruptions occur in seven percent of cases (*I*, 7.2%) and 3.5% concerns semantics (*S*). We should note these are not balanced results - interactions differ to a large degree on both length and times both partners explicitly contribute within each interaction. Figure 2 above offers a clearer view of the distributions.

Table 1. A technical description of the coding as defined

Label	Description
F	<i>Fluent Interaction</i> Turns are appropriately timed and taken, for both question/answers and back-channel cues. Responses are sensible and appropriate
S	<i>Seemingly fluent.</i> Indicates a problem with semantics. Turns are appropriately timed and taken, for both question/answers and back-channel cues. However, responses given are nonsensical and/or out of context.
I	<i>Interruption</i> Wingman takes the turn and interrupts before the utterance is finished or a turn change is indicated.
D	<i>Delay</i> An unnatural delay in the response from the agent after an indicated turn change.
N	<i>No response</i> Virtual wingman does not respond to a question at all, human partner re-takes turn (advances conversation), rephrases or repeats previous statements or verbally indicates it takes too long to respond.

**Fig. 2.** Qualitative data visualized. Values are percentages. **F** = a successful (fluent) interaction; **S** = Seemingly fluent; **I** = Interruption; **D** = Delay; **N** = Nonsensical.

4 Perception Experiment

Handling error conditions in interacting with a virtual agent requires the agent to understand when these errors occur and how they can be identified. To achieve such understanding, we need to look at how conversational partners indicate problems. The perception experiment was aimed at finding evidence that (1) interactions can be correctly recognized given their success and (2) whether or not certain types of miscommunications yield better recognition over others.

4.1 Materials

For selecting the stimuli from the original materials, only three out of seven pilots in the experiment qualified for a balanced set of stimuli with equally divided successful and unsuccessful stimuli. It is therefore not a statistically representative sample from the data in the source materials. Eight (8x1) successful interactions *F* and eight unsuccessful interactions *N, I, S, D* (two from each category, 2x4) were selected from the initial data using random sampling, yielding a total of 48 stimuli (3 pilots x 16 stimuli). Using the stimuli, we made three videos (for three conditions) containing sequences of the 48 stimuli accompanied by a written introduction and questionnaires. We manipulated the videos to create three conditions, differing to the degree of richness in communicative cues: an original condition *AV* with both visual and auditory channels and two lean conditions *A* and *V* with auditory- and visual cues removed respectively.

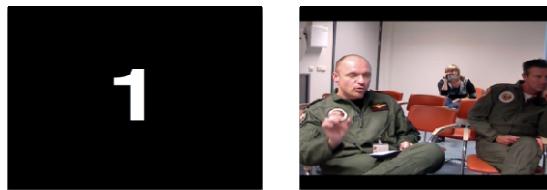


Fig. 3. A screenshot from one of the videos (condition *V*, indicator (1) + a screenshot from the first stimulus on the right).

Sequences in the materials were randomized and balanced by mirroring the sequence to eliminate results in effect of the sequence, yielding a total of six different video compilations (2x3): two different sequences (balance/counterbalance) for each condition (audio, visual, combined).

Questionnaires consisted of a few questions (age, mother tongue, gender) accompanied by a short explanation of the domain and context, a thorough explanation of the task at hand, an instruction and the questions itself. As we wanted to push participants into a decision, a forced choice had to be made: the interaction they were presented with had to be judged in terms of successfulness: each interaction was either a *success* or *no success*. Each question was accompanied by the corresponding number in the video. Participants were seated behind a laptop, and instructed to put on a pair of headphones (for the conditions that contained auditory cues and to block noise for the non-auditory condition) before they randomly received one of six versions of the questionnaires and the corresponding stimuli.

4.2 Participants

A total of 63 students participated in the experiment, 26 males and 37 females aged 18 to 25 ($\mu = 21.84$). Participants were randomly divided over conditions *A/V/AV* in 21/22/20. No participants or cases were excluded from the analyses.

4.3 Design and Analyses

The experiment was set up using a between-subjects design. The analysis was aimed at the (binary) correctness of the answers participants provided. We first recoded participants' answers into a score for recognition of the stimuli based on our pre-scored attribution in a 3x3x2 design (3 pilots by 3 conditions A, V & AV by 2 categories; *success* & *no success*). In the analysis, *Condition* acted as the independent (between-subjects) variable, measuring its effect on *Recognition* (participants' ability to recognize interactions as successful or not). A second analysis, assessing the effect on the recognition values for *Type* of miscommunication (participants' tendency to recognize some miscommunications more easily than others) did not yield a significant influence of either *Type* on recognition scores. Data were analyzed using ANOVA in a General Linear Model with repeated measures in SPSS Statistics.

4.4 Results

As a first indication, a T-test shows participants are able to correctly recognize interactions above chance levels ($t(63) = 8.16$, $P < .001$; $M = 29.16$, $SD = 5.06$) (baseline $\mu_0 = 24$, 48 stimuli). The ANOVA shows there are differences in participants' ability to recognize between successful and unsuccessful interactions ($F(1, 61) = 68.04$, $P < .001$; $\eta^2 = .527$) with success showing significantly higher recognition scores (successful interactions, $M = 18.16$, $SD = 2.81$; unsuccessful interactions, $M = 13.00$, $SD = 5.60$) (*figure 4, left*). There are clear differences in participants' ability to recognize between pilots (some pilots yield better recognition scores) ($F(2, 60) = 22.74$, $P < .001$; $\eta^2 = .431$). On top of that, we see an interaction effect between *Pilot* and *Category* on participants' ability to recognize ($F(2, 60) = 17.89$, $P < .001$; $\eta^2 = .363$), indicating certain types of miscommunications are better observed in certain pilots.

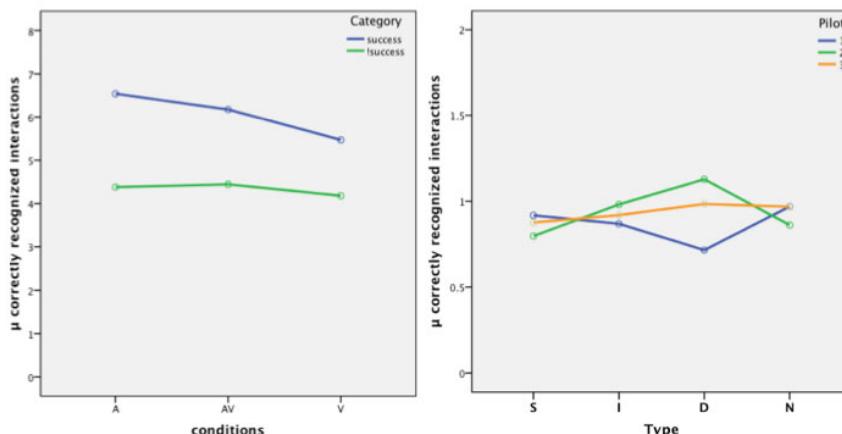


Fig. 4. *Left:* Recognition (Y) scores per category (success | !success) plotted on Condition (X). Y-axis ranges 0 - 8, points represent mean correctly recognized interactions per category per participant for each condition. *Right:* Miscommunications split into Type. (X) plotted for each different pilot. Y-axis ranges 0 - 2, points represent mean correctly recognized interactions per type per pilot. Categories are *Seemingly fluent*, *Interruptions*, *Delays*, *Non-responses*.

There is an interaction effect between *Type* and *Condition*, ($F(6,118) = 3.245$, $P < .001$; $\eta^2 = .042$), indicating differences between conditions on the ability to recognize miscommunications are dependent on *Type* (some miscommunications are more easily observed in certain conditions than others) (figure 4, right). No other significant main- or interaction effects were found.

5 Discussion, Conclusions and Recommendations

5.1 Discussion and Conclusions

The descriptives (in section 3.2) show our initial exploration of what characterizes interaction between a virtual wingman and a human pilot in a setting and context that requires a high degree of expertise (i.e. military aviation and operations). As is apparent, Ashley needs to improve her performance by reducing the number of miscommunications - roughly one in two attempts result in miscommunications. However, more than half of these (one-third of the interactions observed) concern delays, relatively easily avoided by improving conversational timing. Moreover, four out of the five types of miscommunications observed relate to floor management. Thus, when Ashley succeeds in appropriate floor management the number of miscommunications can be reduced by almost half (as much as 45%).

In the perception experiment, we observed participants are able to correctly recognize whether a miscommunication occurred or an interaction was successful. This indicates information on such success can be derived from the interaction between a pilot and a virtual wingman based on cues perceivable in the interactions. It supports the idea that cues provided by a human partner can help optimize the interaction between partners in human-machine dialogue [1] [6]. Interestingly, the mean differences indicate participants are better able to recognize successful interactions than they are to recognize miscommunications. This could indicate problems in the interaction are to a certain degree not perceived as such. Surprisingly, the analysis of the observation data shows an insignificant difference in recognition performance between modalities (audiovisual, visual, auditory). This implicates that in designing the module that will take care of floor management, there is no reason other than e.g. system performance to choose for one modality. Furthermore, the second analysis shows there are no differences in recognition between the types of miscommunications - no one miscommunication is more easily observed.

A first limitation of our study is the availability of data and its variance. The number of pilots of which suitable recordings were available is limited and the available data varied to a high degree. With regard to preparing our data, using a non-naïve perspective and testing it with naïve participants obviously creates a gap - in particular for semantics, where knowledge of the domain is vital to understanding its terminology. For obvious reasons, the coder was not naïve to the domain. Moreover, concerning the coding scheme we chose to focus only on first-level problems within the interactions, disregarding the possibility participants might have judged higher-level problems. As a final limitation, in order to succeed in designing natural dialogue for this context, we should incorporate the human partners' attitude towards flying and working with virtual wingmen, not just the bystanders' perspective we investigated.

5.2 Recommendations

Our study can serve as a first exploration of what happens when pilots engage in conversation with virtual wingmen and where for this domain specifically the development of virtual agents should be heading. We suggest improving the ability to perform appropriate floor management in Ashley and other virtual wingmen to succeed in reducing miscommunications. Generally, if virtual wingmen can sense a change in the state of the conversation and understand the minimal threshold involved after which a silence becomes a turn-change, they can construct immediate responses, informing pilots about their state (“uhmmm, let me think”). On top of improving the experience for the human partner, towards more life-like abilities in natural conversation, it would buy wingmen time to process an utterance or construct a response. Importantly, as our study indicates there is a tolerance towards miscommunications in participants’ perception, it is advised to take a closer look at how this tolerance extends to military pilots. For other miscommunications beside delays, appropriate repair mechanisms should be designed.

In answer to our research questions, it is (1) indeed evident information on the success of an interaction or turn-taking can be derived from the cues pilots elicit, but to a restricted extent. To improve a wingman’s performance, (2) putting effort into capabilities regarding floor management, signaling when turns are given up by the human partner and generating timely responses is advised, as they are the proportionally largest categories of miscommunications that relate to this floor management.

References

- [1] Edlund, J., Gustafson, J., Heldner, M., Hjalmarsson, A.: Towards human-like spoken dialogue systems. *Speech Communication* 50, 630–645 (2008)
- [2] Gratch, J., Rickel, J., Andre, E., Cassell, J., Petajan, E., Badler, N.: Creating interactive virtual humans: some assembly required. *IEEE Intelligent Systems* 17(4), 54–63 (2002)
- [3] Sandercock, J.: Lessons learned for construction of military simulations: A comparison of artificial intelligence to human-controlled agents. DSTOTR-1614, Defence Science and Technology Organisation Systems Sciences Laboratory, Adelaide, South Australia (2004)
- [4] Swartout, W., Gratch, J., Hill, A., Hovy, E., Marsella, S., Rickel, J., Traum, D.: Toward virtual humans. *AI Magazine* 27, 96–108 (2006)
- [5] van Doesburg, W., Looije, R., Melder, W., Neerincx, M.: Face to face interaction with an intelligent virtual agent: The effect on learning tactical picture compilation. In: Prendinger, H., Lester, J.C., Ishizuka, M. (eds.) *IVA 2008. LNCS (LNAI)*, vol. 5208, pp. 490–491. Springer, Heidelberg (2008)
- [6] Martinovsky, B., Traum, D.: The error is the clue: breakdown in human-machin interaction. In: Proceedings of the ISCA Tutorial and Research Workshop Error Handling in Spoken Dialogue Systems. Château d’Oex, Vaud, Switzerland (2003)
- [7] Skantze, G.: Exploring human error handling strategies: implications for spoken dialogue systems. In: Proceedings of ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems, pp. 71–76 (2003)

- [8] Walker, M.A., Litman, D.J., Kamm, A.A., Abella, A.: Evaluating spoken dialogue agents with paradise: Two case studies. *Computer Speech and Language* (1998)
- [9] Dral, J., Heylen, D.K.J., op den Akker, H.J.A.: Detecting uncertainty in spoken dialogues: an explorative research to the automatic detection of a speakers' uncertainty by using prosodic markers. In: *Sentiment analysis: Emotion, Metaphor, Ontology and Terminology*, Marrakech, Morocco, May 27, pp. 72–78. ELRA (2008)
- [10] Cassell, J.: Embodied conversational interface agents. *Communications of the ACM* 43(4), 70–78 (2000)
- [11] Barkhuysen, P., Krahmer, E., Swerts, M.: The interplay between the auditory and visual modality for end-of-utterance detection. *The Journal of the Acoustical Society of America* 123(1), 354–365 (2008)
- [12] Bulyko, I., Kirchhoff, K., Ostendorf, M., Goldberg, J.: Error-correction detection and response generation in a spoken dialogue system. *Speech Communication* 45(3), 271–288 (2005)
- [13] ter Maat, M., Heylen, D.: Turn management or impression management? In: Ruttkay, Z., Kipp, M., Nijholt, A., Vilhjálmsson, H.H. (eds.) *IVA 2009*, vol. 5773, ch. 51, pp. 467–473. Springer, Berlin (2009)
- [14] Swerts, M., Krahmer, E.: Audiovisual prosody and feeling of knowing. *Journal of Memory and Language* 53(1), 81–94 (2005)

A Configuration Method of Visual Media by Using Characters of Audiences for Embodied Sport Cheering

Kentaro Okamoto¹, Michiya Yamamoto¹, and Tomio Watanabe²

¹ 2-1 Gakuen Sanda Hyogo 669-1137, Japan

{kentaro.okamoto,michiya.yamamoto}@kwansei.ac.jp

² 111 Kuboki Soja Okayama 719-1197, Japan

watanabe@cse.oka-pu.ac.jp

Abstract. In sports bars, where people watch live sports on TV, it is not possible to experience the atmosphere of the stadium. In this study, we focus on the importance of embodiment in sport cheering, and we develop a prototype of an embodied cheering support system. A stadium-like atmosphere can be created by arraying crowds of audience characters in a virtual stadium, and users can perceive a sense of unity and excitement by cheering with embodied motions and interacting with the audience characters.

Keywords: Embodied media, embodied interaction, sports cheering.

1 Introduction

In a stadium, spectators can participate in cheering for their team and owing to their presence in large numbers, they can perceive excitement and unity with other supporters of the same team. On the other hand, in sports bars, where people watch sports on TV, it is not possible to experience the atmosphere of the stadium, even though many people may be present in the sports bar. This difference may be attributed to the lack of two factors—sense of presence at the stadium and sense of unity among the spectators and players.

In this study, we focus on the importance of embodiment [1] in sport cheering. Hence, we support embodied sport cheering by configuring visual media as audience by using CG characters and conventional big screens [2] that can provide a sense of reality. Specifically, we focused on actions from the system to the users [3] [4] and on interactions between the system and the users [5] [6]. Then, we developed a prototype of the embodied cheering support system. Here, we create a sense of presence by arraying crowds of audience characters in a virtual stadium, and we create a sense of unity and excitement with embodied motions via actions and interactions between the audience characters and the users.

2 Concept

In this study, we propose a cheering system that provides embodied motions and actions for cheering on the basis of interactions between the system and the users (Fig. 1). There is a large, wide screen in the room, and the system simulates the

stadium using a sports broadcast image and audience characters, thereby creating the sense of presence. The audience characters in the screen excite the users in accordance with the broadcast image, and they generate cheering motions and actions together with the users. In addition, there is a cheering administrator for system control. The administrator controls the system so that anybody can enjoy embodied cheering for the team projected on the screen.

Thus, exciting embodied sport cheering can be realized by watching TV with cheering characters.

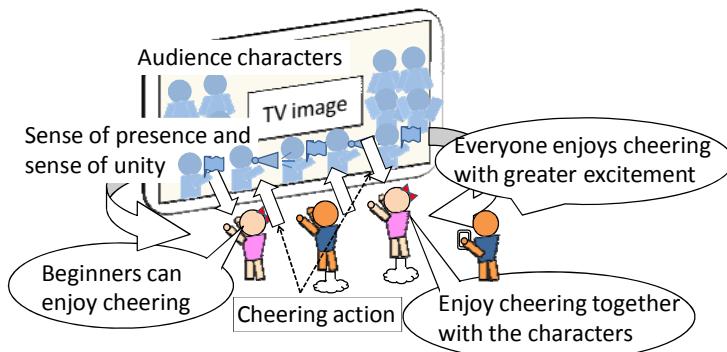


Fig. 1. Concept of the system

3 Embodied Cheering System

3.1 Configuration of the System

Fig. 2 shows the system configuration. A virtual stadium is projected using a large screen (width = 5.8 m, height = 1.8 m), two projectors (EPSON, EB-1735W), a visual computing system (NVIDIA, Quadro Plex 200 D2), and a PC (HP, Z400 Workstation). A soccer game was projected with a BD recorder at the center of the visual stadium by using another projector. We used a speaker for audio output. The iPod Touches, marketed by Apple Inc., was used by the administrator and for inputting users' actions.

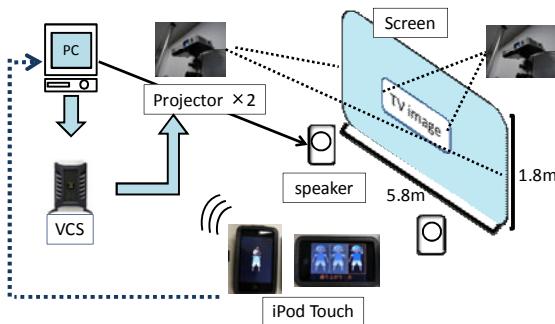


Fig. 2. System configuration

3.2 Configuration of the Virtual Stadium

In order to realize a sense of reality, there are audience seats and characters projected onto the virtual stadium on the large screen. We televised the game at the center of the virtual stadium. Cheerleaders in the audience seats were seated in the front and center rows. The cheerleaders have various cheering tools. There are numerous audience characters in the other seats. Because the television image is incorporated in the virtual stadium, all the users can experience the characters' cheering by simply watching the TV screen (Fig. 3).

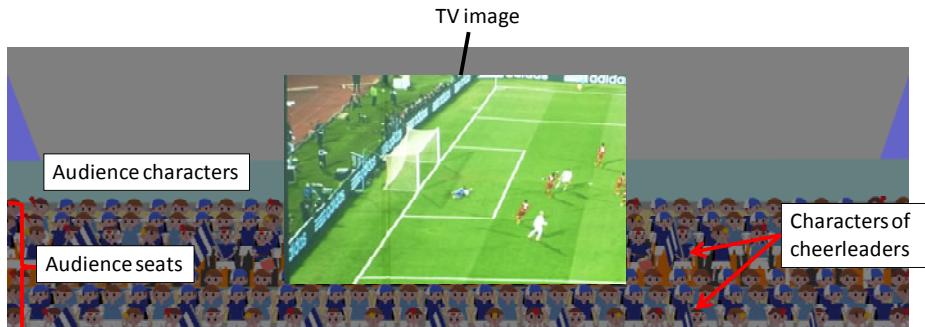


Fig. 3. Configuration of the virtual stadium

3.3 Control of a Character's Action

In this system, we create a sense of presence through actions from the characters in the system to the users, and we create a sense of unity through the interaction between the system and users. Tables 1 and 2 list the characters' actions that create a sense of reality and sense of unity. We explain these actions in more detail later.

Table 1. Sense of presence created by the system

Media	Character	Low	Medium	High
Cheerleaders character	Flag	Slow speed	Average speed	Fast speed
	Vuvuzela			
	Megaphone			
	Drum			

Table 2. Sense of unity through interaction between the system and the users

Media	Character	Low	Medium	High
Audience character	Male	Small jump	Medium jump + Hold up fists	Large jump + Rise arm higher
	Female	Small jump	Medium jump + Hold up fists	Large jump + Waving flag
Vuvuzela	—	Constant volume		

4 Direction of Sense of Reality by the System

4.1 Cheering Action of a Cheer Group

In the virtual stadium, there are many cheering characters waving flags, playing vuvuzelas or drums, and holding megaphones. These characters help to create a sense of presence by simulating the characteristic atmosphere of a stadium through their motions and cheering actions (Fig. 4).

**Fig. 4.** Cheerleader characters

4.2 Control of Cheerleaders by Volume

The motion of a cheerleader character is controlled by the situation of the game. For this purpose, we use the volume of the sports broadcast and the installation site of the system. For example, when the game becomes exciting and the volume of the crowd

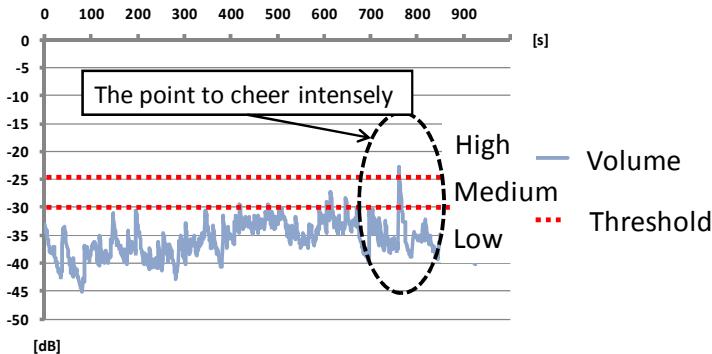


Fig. 5. Control of cheerleaders by using volume

increases, we generate large and rapid motions. Two thresholds control the three levels of motions. Fig. 5 shows an example of the volume and a threshold of the system.

5 Direction of Sense of Unity by Interaction

5.1 Embodied Actions of the Audience Characters

We use the interaction between the users and the audience characters for creating a sense of unity with cheering. Specifically, the motion of the characters depends on the cheering situation, and it is controlled by the system administrator. For example, when the users become excited, the administrator can change the motion of the characters, as shown in Fig 6. To change the motion of the characters, we prepared a level control. Specifically, when the level is low, the characters perform a small jump. The characters smile, raise their arms, and perform big jumps at the middle level. When the level is high, the characters show more joyful expressions than they do at the middle level. Then, the male characters hold up their fists, and the female characters wave a small flag; both male and female characters jump very high.

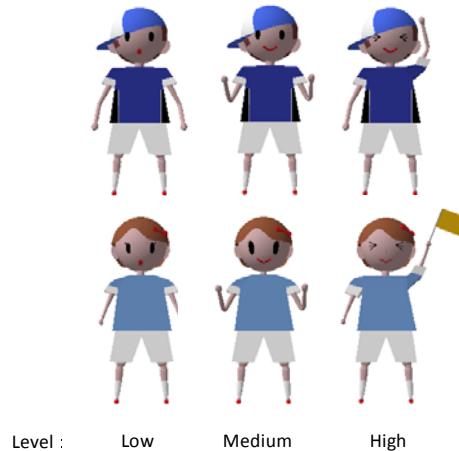


Fig. 6. Jump motions of the characters

5.2 Interaction via Jump

We have measured the acceleration of the users' jump movements to generate the jumping actions of the audience characters. These cheering actions can be shown to many users around the system. The acceleration of the users' movements is measured so that the characters can jump at the same time as the users, providing a level of interaction between the system and the users. Fig. 7 shows an example of acceleration data, as measured by an iPod Touch. We set the threshold for the jump control at the dotted line, and when the acceleration exceeds this value, the audience characters jump (for approximately 0.25~0.4 s).

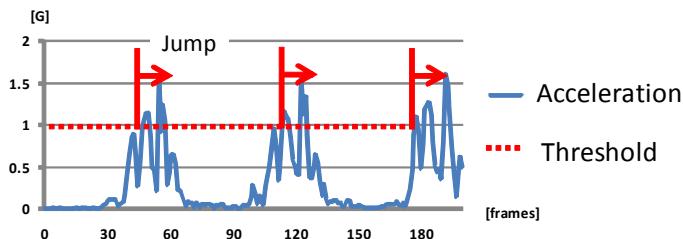


Fig. 7. Jump control by using acceleration

5.3 Interaction via Sound Effects

Users can create vuvuzela sounds by using the iPod Touch (Fig. 8). The sounds indicate excitement and intensify the sense of unity. In addition, the volume of the vuvuzelas intensifies the cheerleaders' actions so that the sense of presence can be enhanced.

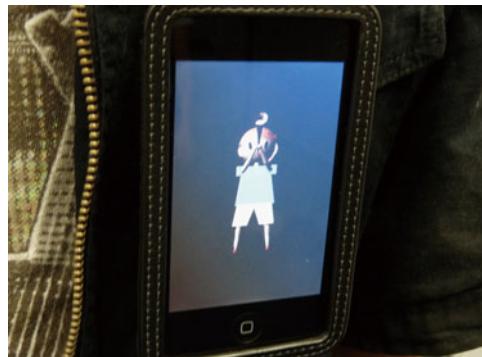


Fig. 8. Screen shot of the iPod Touch

5.4 System Control by an Administrator

We have configured the administrator's iPod Touch so that he or she can control the level of the audience characters' cheering actions (low, middle, or high). Because the

administrator controls the level, users can experience realistic cheering during the game. Furthermore, beginners are able to experience realistic cheering by referring to the characters' actions (Fig. 9).



Fig. 9. Screen shot of iPod Touch for the administrator

6 Using the System

Fig. 10 shows an example of the system usage. Numerous cheerleaders, a virtual stadium, and the televised game are projected on the screen. The users are enjoying the game, and they are experiencing a sense of presence and unity by supporting a team via embodied cheering.

Users have commented that they could enjoy cheering together with the characters because the characters jump and hold their fists according to the users. Some users mentioned that the experience was more enjoyable than the usual television broadcast. Furthermore, some users believe that the system should be provided with more functions to promote cheering. Thus, we confirmed the effectiveness of the system.



Fig. 10. Example of the system

7 Conclusion

In this paper, we proposed a configuration of visual media that promotes sport cheering via physical interactions with many virtual characters, and we developed a prototype system. In addition, we confirmed the effectiveness of the system.

Acknowledgements. This work under our project “Embodied Communication Interface for Mind Connection” has been supported by “New IT Infrastructure for the Information-explosion Era” of MEXT KAKENHI. Also, our project “Generation and Control Technology of Human-entrained Embodied Media” has been supported by CREST of JST.

References

1. Tomitsch, M., Aigner, W., Grechenig, T.: A Concept to Support Seamless Spectator Participation in Sports Events Based on Wearable Motion Sensors. In: The Second International Conference on Pervasive Computing and Applications (ICPCA 2007), pp. 209–214 (2007)
2. Nagahara, H., Yagi, Y., Yachida, M.: Super-resolution Modeling and Visualization System with Wide Field of View. In: SICE 2003 Annual Conference, vol. 3, pp. 3224–3229 (2003)
3. Watanabe, T., Okubo, M., Nakashige, M., Danbara, R.: Interactor: Speech-driven embodied interactive actor. International Journal of Human-Computer Interaction 17, 43–60 (2004)
4. Ishii, Y., Osaki, K., Watanabe, T., Ban, Y.: Evaluation of embodied avatar manipulation based on talker’s hand motion by using 3D trackball. In: Robot and Human Interactive Communication, pp. 653–658 (2008)
5. Sejima, Y., Watanabe, T.: An embodied virtual communication system with a speech-driven embodied entrainment picture. In: Robot and Human Interactive Communication, pp. 979–984 (2009)
6. Sejima, Y., Watanabe, T., Yamamoto, M.: Analysis by Synthesis of Embodied Communication via VirtualActor with a Nodding Response Model. In: Proc. of Second International Symposium on Universal Communication (ISUC 2008), pp. 225–230 (2008)

Introducing Animatronics to HCI: Extending Reality-Based Interaction

G. Michael Poor¹ and Robert J.K. Jacob²

¹ Computer Science, Bowling Green State University

Bowling Green, OH, 43403

² Computer Science, Tufts University

Medford, MA 02155

gmp@bgsu.edu, jacob@cs.tufts.edu

Abstract. As both software and hardware technologies have been improved during the past two decades, a number of interfaces have been developed by HCI-researchers. As these researchers began to explore the next generation of interaction styles, it was inevitable that they use a lifelike robot (or animatronic)-as the basis for interaction. However, the main use up to this point for animatronic technology had been “edutainment.” Only recently was animatronic-technology even considered for use as an interaction style. In this research, various interaction styles (conventional GUI, AR, 3D graphics, and introducing an animatronic user interface) were used to instruct users on a 3D construction task which was constant across the various styles. From this experiment the placement, if any, of animatronic technology in the reality-based interaction framework will become more apparent.

Keywords: Usability, Animatronics, Lifelike Robotics, Reality-Based Interaction, Interaction Styles.

1 Introduction

Since the early 1980s, the graphical user interface (GUI) has been the defacto user interface (UI) associated with computing. No matter which platform – Windows, Mac OS, or Linux – when people use a modern computer, as well as most cell phone and handheld devices, they are interacting with a GUI. However, because of the advances in computing technology in both hardware and software during the past few years, there has been increased interest in designing the next “standard” UI. From this interest, researchers have begun developing a broad range of new interaction techniques that split from the “window, icon, menu, pointing device” (WIMP) interaction style that is prevalent today [1]. These post-WIMP interaction styles have been defined by van Dam as “containing at least one interaction technique not dependent on classical 2D widgets such as menus and icons” [2]. Some examples of these post-WIMP interaction styles are tangible user interfaces, as defined by Ishii and Ullmer [3], context-aware interfaces, as defined by Schilit et al. [4]; and a number of other unique systems [5, 6, 7].

Nevertheless, no one next-generation UI has been labeled the next default UI. Researchers at Tufts University have observed that these next-generation interaction styles include many similarities between them [1]; they all appear to use interaction techniques that incorporate reality-based movements, actions, and concepts. User's knowledge gained from experiences in the real world and people's interactions with objects can be incorporated in such a way that people will interact with the next generation of computers in a more natural and less mentally taxing context. This idea led to the framework called reality-based interaction (RBI). According to RBI, when an interaction style incorporates actions based on preexisting knowledge of the nondigital, everyday world, it is easier for users to learn to interact with a new system. The central claim of RBI is that the more people have used or encountered real-world phenomena, the easier it is for the users to call on that knowledge.

RBI theorists have attempted to take the next step in including and characterizing additional kinds of input as well as identifying the trend in the realm of human-computer interaction (HCI) as heading toward more reality-based interaction styles. This study examines whether manipulating the RBI characteristics of an interface has an effect on a user's performance. More specifically, is there a difference in the overall performance of a user's ability to complete a construction task delivered using four interaction styles with varying levels of reality-based interaction techniques?

By reviewing the four themes of reality-based interaction as identified in the RBI framework and identifying potential future interaction styles that would incorporate these four themes, the topic of animatronics and the potential use as an interaction style has become a major subject of interest. Animatronics exhibited many qualities that are reality-based in nature and a number of design process goals were identified:

1. Determine whether an animatronic interface could be used as a reality-based interaction style.
2. Determine how well the animatronic interface compares to other established interaction styles.

In this paper previous work related to the RBI framework and its history are discussed as well as an investigation of the current state of animatronics and the issues pertaining to the technology. This is followed by the description of the experiment; including condition design/creation and addition experimental details. Finally a review of the results is presented in addition to a discussion.

2 Related Works

In this experiment, there are three different types of interaction that the subjects could potentially use; each of which has their own history and bodies of research to explore. Highlights about each type follow.

2.1 Reality-Based Interaction

With the improvements in both software and hardware technology over the past few decades, there has been an increase in the number of new interfaces that have been developed by human-computer interaction researchers. At a cursory glance, these new interaction styles might not appear to have much in common. It has been proposed by

Jacob et al. [1] that “[these new interaction styles] share salient and important commonalities, which can help us understand, connect, and analyze them” (p. 1). One of the biggest commonalities is the increased ability to draw upon the users’ preexisting knowledge of the real world, which is a trait that is present throughout all of these new interaction styles. By investigating these commonalities and the various changes that these interaction styles have gone through during their development, RBI would provide researchers with a “lens” through which they could gain insights for design and discover new avenues for research.

When a user attempts to learn a new system, a number of problems must be confronted. According to Norman et al. [8], two important problems will be faced: the “gulf of execution” and the “gulf of evaluation.” The gulf of execution is the mental gulf that users must cross to turn their intentions into commands so that the interface can perform a task. The gulf of evaluation is the mental gulf that users must cross to understand the state of the system that results from the interface’s feedback (after a task has been executed). According to Jacob et al., when four themes of reality are incorporated, emerging interaction styles attempt to bridge the gap between these two problems. The four themes are:

- *Naive Physics*: the concept that people have commonsense knowledge about the physical world.
- *Body Awareness*: people's awareness of their own physical bodies and their skills for controlling and coordinating their bodies.
- *Environment Awareness*: people's sense of their surroundings and their skills for negotiating, manipulating, and navigating within their environment.
- *Social Awareness*: people generally are aware of others in their environment and have skills for interacting with them.

The bridge that is created by these themes is the result of the users’ familiarity with these themes and the ways in which they perform interactions that incorporate them without additional effort. The lack of required interpretation of intentions into the interface’s language frees the user to perform the interactions automatically [10]. This type of automation also can be applied to the translation of the system’s feedback, which allows the user’s preexisting knowledge of the real world to be the predominant evaluation tool [9].

Incorporating interfaces with reality-based interaction characteristics so they completely mimic reality is not an optimal solution. According to this framework, RBI principals are important to consider during development, but there are times when this is not practical. Jacob et al. proposed that if a researcher gives up reality-based interaction, then it should be given up “explicitly and only in return for other desired qualities” (p. 5). Jacob et al. defined these desired qualities as:

- *Expressive Power*: users can perform a variety of tasks within the application domain.
- *Efficiency*: users can perform a task rapidly.
- *Versatility*: users can perform many tasks from different application domains.
- *Ergonomics*: users can perform a task without physical injury or fatigue.
- *Accessibility*: users with a variety of abilities can perform a task.
- *Practicality*: users find the system is practical to develop and produce.

It is through the “lens” of RBI that researchers have been able to analyze and compare various designs, bridge an assortment of problem gaps between seemingly unrelated research areas, and apply the lessons learned in the development of one interaction style to another style.

2.2 Research in Animatronics

Currently, common themes in animatronics research have robots learn from, work collaboratively with, or assist human users [11]. In the area of human-robot partnership, Cynthia Breazeal is attempting to escape from the traditional artificial intelligence goal of creating human-equivalent intelligence in technological systems. Instead she is attempting to create robots that bring value to a human-robot relationship in such a way that humans can appreciate robots for the ways that they enhance our lives and complement people's strengths and abilities. The focus of this avenue of research has been on either the human subject teaching the robot [12] or the robot and human working as a team [13]. The two-way dynamic has not been the animatronic being used as an interactive system to instruct the human subjects.

3 The Study

In order to systematically investigate whether the amount of reality-based interaction has an effect on a subject's ability to interact with a system, four interaction styles were developed, with varying levels of reality-based characteristics. Each of these conditions concerned the same task and the interactions were consistent. Additionally, the number of reality-based attributes used for each condition was augmented for each reality-based interaction style.

Four reality-based attributes were identified as being augmentable:

- *Motion*: utilizing or not utilizing a full range of motion, from static pictures to dynamic movement of the avatar, to the instruction delivery method.
- *Resolution*: the clarity with which the user can see the details of the instruction delivery method ranging from reality (no computer assistance), to 3D graphical rendering, to digital photos viewed on a computer monitor.
- *2D vs. 3D*: the presentation of the information can be varied depending on the instruction delivery method.
- *User-Controlled Depth of Field*: giving the user the ability to change their perspective of the instruction delivery method.

By selecting varying levels of these attributes, we were able to identify four varying levels of reality-based interaction that are used in this experiment. The GUI condition was chosen as the default interaction style because it is currently the most widely used interaction style and hence provides a good baseline. The decision to use the 3D condition was based on the trend in research to use virtual representations of avatars to interact with subjects. The AR condition was then included as an interaction style that had similar properties to the 3D condition, however it included additional reality-based characteristics that were not present in the 3D condition.

3.1 Condition Design

After a series of pilot tests were run using human actors as the instruction delivery system it was concluded that the animatronic condition could incorporate similar interactions. The pneumatically controlled animatronic character (Figure 3) was designed to be in a sitting position with moves that are considered the typical moves for a sitting museum-quality animatronic figure, providing enough body and facial movement to be considered realistic. In order to give the illusion of interactivity for the subject, the “Wizard of Oz” (WoZ) technique was implemented.

In the creation of the GUI condition (Figure 1), it was concluded that using the ZOOB pieces was not an equivalent task to the other three conditions. There was far more information about the specific step being supplied with their inclusion. By taking still images of the animatronic character throughout various points in its movements the interaction that was consistent with the other three conditions. For each step, at least a beginning position and an end position were shown in a comic-book-style format. If the move was so complex that it required additional information, a third picture was included that displayed the position of the character mid-move.

Finally, a 3D representation of the animatronic character was created that when viewed through a head-mounted display would give the appearance of a 3D computer-generated character in the real world for the AR condition. That same character would also be used in a desktop setting for the 3D condition (Figure 2).



Fig. 1. GUI Condition



Fig. 2. 3D and AR Conditions



Fig. 3. Animatronic Condition

4 Experiment

The study was a between subjects design conducted with 80 undergraduate students randomly assigned, between the ages of 18 and 31, at Tufts University.

Each step was delivered to the subjects via their specific interaction style. To start, the subjects were given all of the individual ZOOB pieces that they would need. There were no distracter pieces included, so the subject would include all supplied pieces in the construction task when it was completed. The instructions were delivered to the subjects one step at a time, giving them as much or as little time as they required to complete the actions.

During step delivery, the subjects were allowed to interact with the pieces however they saw fit. Each step was either a legal connection between pieces that was inherent in the ZOOB, a rotation of a ZOOB piece, a tilt or angling of a ZOOB piece, or some combination of those three. The subsequent step was not provided until the subjects indicated that they had completed the step by placing the semi-finished construction on the yellow “X” and that step had been deemed correct by the researcher. Through the illusion created by the WoZ effect, it would appear to the subject that the system was reacting to their choices.

If a step was completed incorrectly, the interface would supply the subject with the corresponding error instruction, and then it would repeat the incorrect step so that the subject would attempt to perform the step correctly. The subject could attempt to correct the error at any time. An incorrect connection, piece, tilt, or rotation could hinder subsequent steps, so the error messages were continued until either the current step was performed correctly or the researcher deemed the error was “close enough.” This means that the step that was performed was correct in regards to the information that was supplied, but might not be correct in terms of the precise step. The alternative option was that the subject could require manual assistance from the researcher.

4.1 Experimental Design

This study employed one independent variable, a covariate (spatial perception), and seven dependent variables. The independent variable was the varying levels of reality-based interfaces. The four different levels of reality-based interfaces included (1) the

slideshow on a 2D monitor, (2) the 3D instructions delivered on a 2D monitor, (3) the 3D instructions delivered in an AR environment by an HMD, and (4) the 3D instructions delivered by an animatronic character. A behind-the-scenes person was in control of all the conditions, employing the WoZ technique. The WoZ technique directs the conditions to deliver each of the steps to the subjects, giving the illusion that the computer is adjusting its reactions to the subject's actions. As with all of the conditions, the subjects were not given any way of controlling the interface other than completing each step correctly.

For this experiment, performance was measured in terms of seven dependent variables. These dependent variables were divided into two groups for no other purpose but for ease of presentation. The first group included the number of times a step was repeated, the number of times human assistance was requested, and the total number of errors. The second group included the four types of errors that a subject could perform: incorrect connection, incorrect rotation, incorrect piece selection, and incorrect tilt or angle. A higher occurrence of these dependent variables means a lower performance by the subjects.

5 Results

Once all of the data was collected, it was found that there was a significant difference between interaction styles in all three of the first group's dependent variables. There were also significant differences between interaction styles in two of the dependent variables from the second group, incorrect piece and incorrect angle errors. Thus we are able to say that there is a significant difference in performance of a user's ability to complete a construction task delivered by four interaction styles with varying levels of reality-based interaction techniques. The specifics of the results follow.

The first test performed was a multiple analysis of covariance (MANCOVA) to investigate the effects of the independent variable and the covariate, condition and spatial ability, on a subject's performance as measured by the seven dependent variables. The results of the MANCOVA indicate that the covariate, spatial ability, relates significantly with the dependent variables (Wilks' Lambda, $F = 4.388\$; p < 0.001\$$). Controlling for spatial ability, the effect of the independent variable, interaction style, was still significant for the dependent variables (Wilks' Lambda, $F = 2.173\$; df = 3\$; p < 0.005\$$). Once the MANCOVA was completed, an analysis of covariance (ANCOVA) was conducted on all seven of the dependent variables, taking into consideration both independent variables.

According to the results of the ANCOVA, a significant difference was found between all of the dependent variables with regards to spatial ability except incorrect piece errors and incorrect angle errors. When controlling for spatial ability, the results of the ANCOVA indicate that the effects of interaction style cause significant differences in the first three dependent variables: repeats ($sig = 0.012\$$), outside help ($sig = 0.020\$$), and errors ($sig = 0.042\$$). Of the remaining four dependent variables, only the incorrect rotation error ($sig = 0.026\$$) showed significance when we controlled for spatial ability and took interaction style into consideration.

Given the significant difference between conditions for number of dependent variables when controlling for spatial perception, a multivariate analysis of variance

(MANOVA) was performed to investigate the effect of condition on the seven dependent variables. The results of the MANOVA were significant (Wilks' Lambda, $F(3, 75) = 2.457\$; p = 0.001\$$), signifying that across the seven dependent measures, the independent variable of varying levels of reality had a significant effect. This means a between-subjects analysis of variance (ANOVA) can be done.

The ANOVA was performed on all seven of the dependent variables. This test determined whether the independent variable, varying levels of reality-based interaction styles, had any impact on each specific dependent variable. According to the results of the ANOVA, the independent variable had a significant impact on the number of repeat requests ($F(3, 75) = 5.064\$; p = 0.003\$$), the number of outside help requests ($F(3, 75) = 4.237\$; p = 0.008\$$), the number of errors committed ($F(3, 75) = 4.640\$; p = 0.005\$$), the number of incorrect piece errors ($F(3, 75) = 3.909\$; p = 0.012\$$), and the number of incorrect angle errors ($F(3, 75) = 4.577\$; p = 0.005\$$). Thus there is a significant difference between the four varying levels of reality-based interaction styles in terms of the aforementioned dependent variables. It must be noted, however, that the results of the ANOVA and the ANCOVA are not completely consistent in terms of the types of errors that were performed. It must be noted that the effect that was observed in that portion of the ANOVA could be due to the effect of the subject's spatial ability and the results have to be viewed while taking that into consideration.

6 Discussion

As was shown in the statistical results, there was a significant difference between interaction styles in all three of the first group's dependent variables: steps repeated, assistance requests, errors. There were also significant differences between interaction styles in two of the dependent variables from the second group: incorrect piece and incorrect angle errors. Therefore, given these results, we are able say that there is a significant difference in performance of a user's ability to complete the same construction task delivered by four interaction styles with varying levels of reality-based interaction techniques.

When delving further into the analysis of the overall experiment, there is evidence that users in the AR condition consistently performed much worse than those in other conditions. In both number of repeats and number of errors, the AR condition was significantly different than all of the other conditions. The outside help dependent variable also had a significant difference between the AR condition and the GUI condition. However, the other two conditions, although not significant, were approaching significance when compared to the AR condition.

When this experiment commenced, a formal ranking was not applied to the conditions; however, an informal ranking was conceived. It was concluded that the results would inform the researchers of the ranking after the experiment was completed. The resulting rankings did not coincide with the initial intuition of the researchers. It was believed that the GUI condition would perform the worst, followed by the 3D condition, then the AR condition, with the animatronic condition performing the best. These initial rankings stemmed from extensive dialogue with consultants and experts in HCI at Tufts, as well as experts in HCI and statistics at

Bowling Green State. The logic behind the original intuition was due to the apparent reality-based characteristics that are present in the technologies and varying amounts of the four themes identified in the RBI framework.

However, when realistic budgets and constraints were placed on the technologies, the quality of these technologies and their overall ability was hindered. The rankings that were conceived from the RBI framework at the outset of this experiment did not foresee the technological inhibitors that would be present due to the state of the art in animatronics, and the budget and time constraints. Even with budget, time, and technological inhibitors, the animatronic condition was able to perform at the same level as the top-performing technologies and statistically better than the AR condition.

7 Conclusion

We have presented a study of four interaction styles with varying levels of reality-based interaction techniques each being used to convey instruction for a 3D construction task. It was found that three of the conditions (GUI, 3D, and Animatonic) all performed in a roughly equivalent manner. However, the AR condition was found to perform statistically worse in regards to the dependent variables. In terms of the RBI framework, this was not the expected results. The higher levels of “reality-based” characteristics should have caused a different ranking in regards to performance.

Now why did this happen? One thought is that perhaps the interactions required with AR conditions were too overwhelming for the subjects. In other words, the subjects were perhaps not able to conceive the scope of interaction that was possible in the condition and became intimidated, consequently hindering their performance. Another thought was that even though the AR condition, by its nature, attempted to draw from real-world interactions, the uniqueness of the technology might have made that type of experience almost impossible, given the time that the subjects had to interact with it and become acclimated to it.

The other important point to note is that the animatronic condition performed as well as two interaction styles (3D and GUI) that are widely accepted and in everyday use. It was concluded that given advancement in technology and an increase in budget, the animatronic condition would be able to achieve the higher performance levels suggested by the RBI framework, and has the potential to perform better than these technologies in this type of task. The needed advancements in the technology are still a few years away, but animatronics will become an integral part of the growing robotic world.

References

1. Jacob, R.J., Girouard, A., Hirsheld, L.M., Horn, M.S., Shaer, O., Solovey, E.T., Zigelbaum, J.: Reality-based interaction: A framework for post-WIMP interfaces. In: CHI 2008: Proceeding of the Twenty-sixth Annual SIGCHI Conference on Human Factors in Computing Systems, pp. 201–210. ACM, New York (2008)
2. van Dam, A.: Post-WIMP user interfaces. Communications of the ACM 40, 63–67 (1997)

3. Ishii, H., Ullmer, B.: Tangible bits: towards seamless interfaces between people, bits and atoms. In: CHI 1997: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 234–241. ACM, New York (1997)
4. Schilit, B., Adams, N., Want, R.: Context-aware computing applications. In: Proceedings of Workshop on Mobile Computing Systems and Applications, pp. 85–90 (1994)
5. Boud, A.C., Baber, C., Steiner, S.J.: Virtual reality: A tool for assembly? *Presence: Teleoperators & Virtual Environments* 9, 486–496 (2000)
6. Iqbal, R., Sturm, J., Kulyk, O., Wang, J., Terken, J.: User-centred design and evaluation of ubiquitous services. In: SIGDOC 2005: Proceedings of the 23rd Annual International Conference on Design of Communication, pp. 138–145. ACM, New York (2005)
7. Tang, A., Owen, C., Biocca, F., Mou, W.: Comparative effectiveness of augmented reality in object assembly. In: CHI 2003: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 73–80. ACM, New York (2003)
8. Norman, D.A., Draper, S.W.: User Centered System Design: New Perspectives on Human-Computer Interaction. Lawrence Erlbaum Associates, Inc., Mahwah (1986)
9. Christou, G.: A Knowledge-Based Framework for the Description and Evaluation of Reality-Based Interaction, Ph.D. thesis, Tufts University (2007)
10. Logan, G.: Toward an instance theory of automatization. *Psychological Review* 95, 492–527 (1988)
11. Coradeschi, S., Ishiguro, H., Asada, M., Shapiro, S., Thielscher, M., Breazeal, C., Mataric, M., Ishida, H.: Human-inspired robots. *IEEE Intelligent Systems* 21, 74–85 (2006)
12. Lockerd, A., Breazeal, C.: Tutelage and socially guided robot learning. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2004, vol. 4 (2004)
13. Hoffman, G., Breazeal, C.: Effects of anticipatory action on human-robot teamwork efficiency, fluency, and perception of team. In: Proceedings of 24 the ACM/IEEE International Conference on Human-robot Interaction, pp. 1–8. ACM, New York

Development of Embodied Visual Effects Which Expand the Presentation Motion of Emphasis and Indication

Yuya Takao¹, Michiya Yamamoto¹, and Tomio Watanabe²

¹ 2-1 Gakuen Sanda Hyogo 669-1137, Japan

{yuya.takao,michiya.yamamoto}@kwansei.ac.jp

² 111 Kuboki Soja Okayama 719-1197, Japan

watanabe@cse.oka-pu.ac.jp

Abstract. Although visual presentation software typically has a pen function, it tends to remain unused by most users. In this paper, we propose a concept of embodied visual effects that expresses emphasis and indication of presentation motions using a pen display. First, we measured the timing of presentation motions of pen use achieved while in sitting and standing positions. Next, we evaluated the timing of underlining and explanation through a synthesis analysis from the viewpoint of the attendees. Then, on the basis of the results of our measurements and evaluation, we developed several visual effects. These visual effects, which express the embodied motions and control the embodied timing, are implemented as system prototypes.

Keywords: Human interaction, presentation, timing control, visual effect.

1 Introduction

Recently, the number of situations requiring the use of information devices in presentations is increasing. Presentation software and video have become particularly popular elements of visual presentations. In addition, some researchers supports presenter's gesture, gaze and other forms of paralanguage by displaying presenter's figure in slides or by enabling presenters to control presentation system with bare hand [1],[2], or the other researchers have developed presentation support system by building more useful function with Pen display [3]. However, presentations using information devices can decrease the presenter's embodied motions and actions. Moreover, although presentation software generally includes a pen function, most users seem to avoid using it. One reason for this may be a difficulty for users to make themselves understood via embodied motions and actions while using information devices. In addition, the timing of the drawing and other pen effects are generally awful [4].

By focusing on the entrainment of embodied rhythms, the author in [5] has promoted the study of essential human interaction. As an example, the authors have analyzed greeting interactions between a human and a robot and between a human and a CG character, which made clear that a difference in timing to generate communicative actions and utterances can alter the effects of communication [6]. On the basis of these results, it will become possible to support communication to audiences by generating visual effects and delay the timing.

In this paper, we support a presentation that uses information devices by introducing the presenter's embodied motion and actions into a system via a pen display. We then control the drawing timing using visual effects. To achieve this, we first analyzed the drawing timing of a presenter accustomed to the pen function and then conducted an evaluation of the audience response. Next, we developed the embodied visual effects.

2 Concept

There are several approaches to support visual presentations using information devices. For instance, various wireless presenters are available in the market, and many researchers have made progress in the area of visual effects for presentations using laser pointers [7],[8],[9],[10]. However, using a laser pointer is useless in a large venue as the audience cannot see the presenter's motions and actions, and the pointer controlled by the presenter may be blurred on the screen.

In this study, we use a pen display to enhance the presenter's embodied motions and actions with the pen function. We chose a pen display as it is intuitively suitable for gesture inputs and expanding the embodiment. For this purpose, when we use a pen function to emphasize a part of a slide, visual effects are displayed on the slide. In this study, we propose two kinds of effects, one is for introducing the presenter's actions and the other is for controlling the timing when emphasizing certain aspects of a slide (Fig. 1).

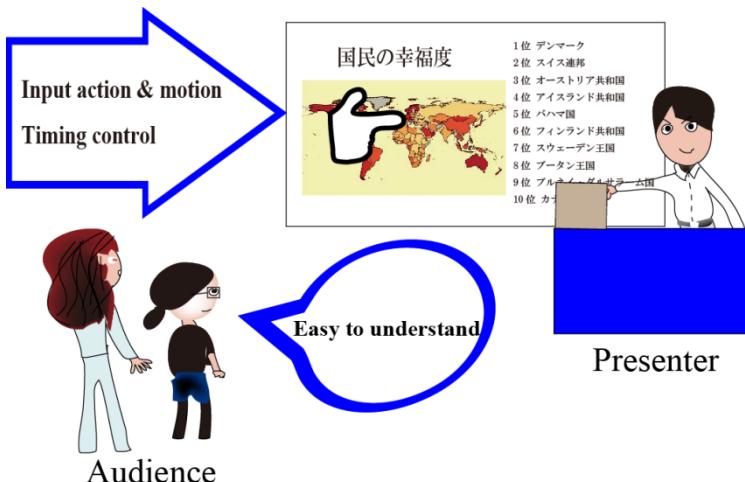


Fig. 1. Concept of the embodied presentation system

3 Drawing Experiment by a Presenter

3.1 Experimental Setup

In our previous study, we made it clear that the timing needed to use a pen function as a pointer or for making utterance during presentations varies significantly and that the

timing is difficult to control equally. However, this was because the subjects used in the study were not accustomed to using the pen display. Therefore, we analyzed the presenter's motions and actions during a presentation by assuming that the presenter typically uses a pen display instrumentally. Figs. 2 and 3 show the system configuration. In each presentation, the slides are displayed on a pen display (WACOM, DTI-520), and the presenter explains and underlines the slides on a display. A PC records the utterances by using microphone (ELECOM, HS-HP03BK) and timing of underlining motions. There were six slides used (Microsoft, PowerPoint 2007) in the presentation, and the words were underlined in red.

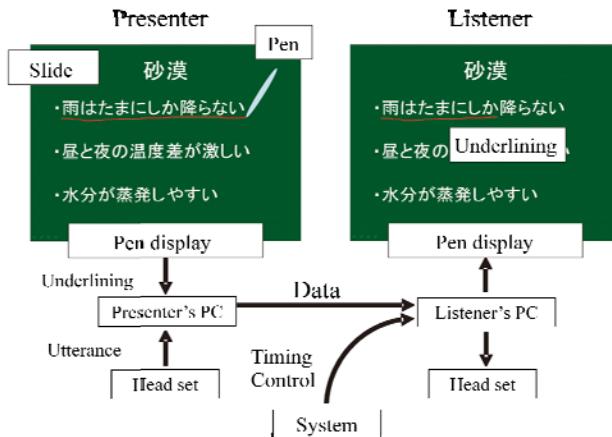


Fig. 2. System configuration for a sitting position

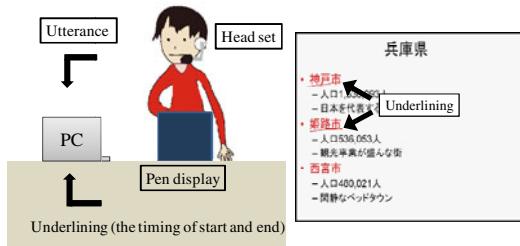


Fig. 3. System configuration for a standing position

During the experiment, the presenters repeated the same slides for 15 min, after which they took a 5-min break before repeating the presentation again for an additional 15 min. Each experiment was performed by 10 Japanese students, 20–24 years in age. The subjects each gave presentations for an audience in both sitting and standing positions. While the subjects gave their presentations in a sitting position, the system sent the presentation data to a listener in another room (Fig. 4). For their standing presentations, we set up a video camera (SONY, HVR-1500A) to record the subjects (Fig. 5).



Fig. 4. Experimental set up for a sitting position



Fig. 5. Experimental set up for a standing position

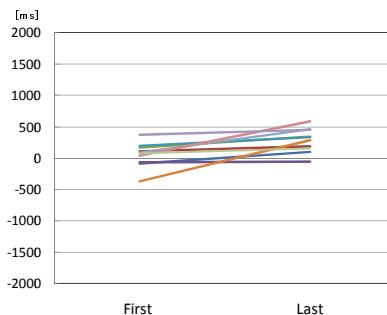


Fig. 6. Results obtained during a sitting position

3.2 Results

Fig. 6 shows the difference in timing between an utterance and drawing motion during a sitting presentation. The data indicates the average time for the first two and last two presentations. When the timing of drawing is faster than that of utterance, the data shows a positive value. As a result, the values obtained from 7 of the 10 subjects are positive during the first half, while the values obtained from 9 of the 10 subjects are positive during the last half. Next, Fig. 7 shows the difference in timing between an utterance and the drawing motion during a standing presentation. As a result, the values obtained from 9 of the 10 subjects are negative during the last half. However, based on video confirmation, it took 0.5–1.5 s from the start of a drawing motion to the beginning of the actual underlining. Fig. 8 shows the difference in timing between

an utterance and the start of a drawing motion during a standing presentation. We found that the timing of the underlining motion occurs about 0.5 s later than the utterance and that the timing of the beginning of the drawing motion is 0–1.0 s faster than the utterance. Fig. 9 illustrates these relationships. Namely, both standing and sitting presentations show similar trends in their drawing timing.

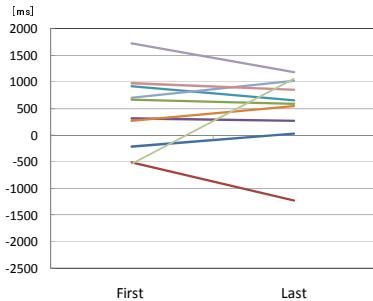


Fig. 7. Start of an underlining motion

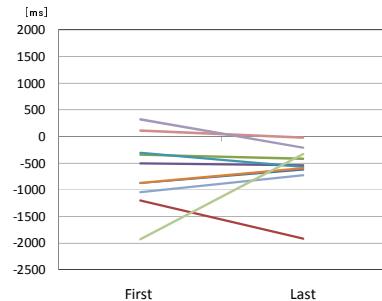


Fig. 8. Start of a motion

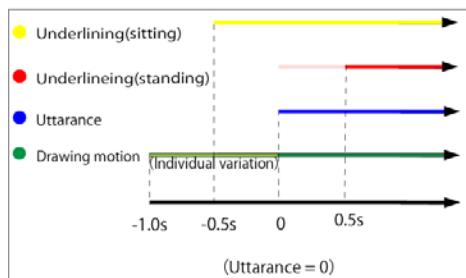


Fig. 9. Results obtained during a sitting position

4 Evaluation Experiment by Audience

4.1 Experimental Setup

To clarify the preferred timing of an utterance and underlining motion for audiences during a presentation, we performed an evaluation experiment based on the results presented in Chapter 3. We referred to two types of timings. The first is the data for when a drawing motion is 270 ms faster than an utterance, while the other is the data for when the drawing motion is 280 ms later than an utterance. There were three kinds of patterns compared: (a) an underlining motion set 270 ms earlier than an utterance, (b) no control used, and (c) an underlining motion set 280 ms slower than an utterance. During the experiment, patterns (a) through (c) were subjected to a paired comparison for both the types of timing data. Two of the six patterns ($2 \text{ data} \times 3 \text{ patterns}$) were selected randomly. The subjects were instructed to select their “preferred” pattern between the two. The experiment was given to 12 (6 male and 6 female) Japanese students.

4.2 Result

Here, the Bradley-Terry model was fitted to the results for a quantitative analysis, as shown in the following equations.

$$P_{ij} = \frac{\pi_i}{\pi_i + \pi_j}, \text{ and} \quad (1)$$

$$\sum_i \pi_i = \text{const.} (=100), \quad (2)$$

where π_i is the intensity of preference for pattern i , and P_{ij} is the probability that i is judged better than j .

The left-hand side of Table 1 shows the results of the paired comparisons for data on the drawing motion 270 ms faster than an utterance. Based on the Bradley-Terry model, preference π was estimated as shown in the left-hand side of Fig. 10, where (c) was rated twice that of (b) and (a), which were almost equal. The right-hand side of Table 1 shows the results of the paired comparisons for data on the drawing motion 280 ms slower than an utterance. On the basis of the Bradley-Terry model, preference π was estimated as shown in the right-hand side of Fig. 10, where (a) is rated more than twice as high as (b), which was rated higher than (c).

From these results, the pattern in which the underlining motion and utterance are generated at the same time was preferred by most of the audience members.

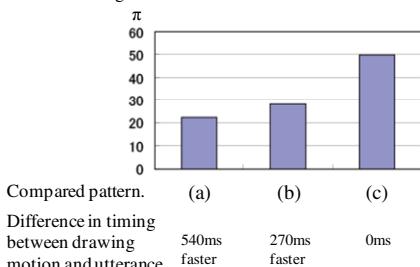
Table 1. Results of a paired comparison

Underlining motion 270 ms faster than an utterance. Underlining motion 280 ms slower than an utterance.

	Faster	0	Slower	Total
Faster		5	4	9
0	7		4	11
Slower	8	8		16

	Faster	0	Slower	Total
Faster		9	9	18
0	3		8	11
Slower	3	4		7

Underlining motion 270 ms faster than an utterance.



Underlining motion 280 ms slower than an utterance.

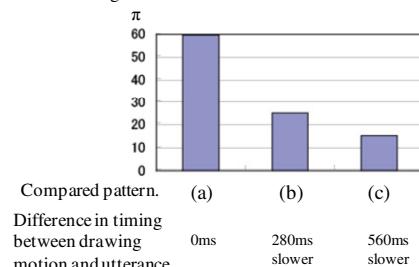


Fig. 10. Comparison of π based on the BT model

5 Embodied Visual Effects

5.1 Configuration of the System

On the basis of the results presented in Chapters 3 and 4, we found that an underlining motion is faster or slower than an utterance during presentations with a pen, but when the timing of an underlining motion and utterance are the same, audience reaction was rated high. Therefore, we have developed visual effects that can fill in the gap by changing the presenter's timing to the actual motion and adjusting it to the utterance. Fig. 11 shows the system configuration. This system is composed of a pen display (WACOM, DTI-520), a PC, a projector, and a screen. In the presentation software, visual effects are shown on slides (Microsoft, PowerPoint 2007), and the slides are shown on a pen display and a screen.

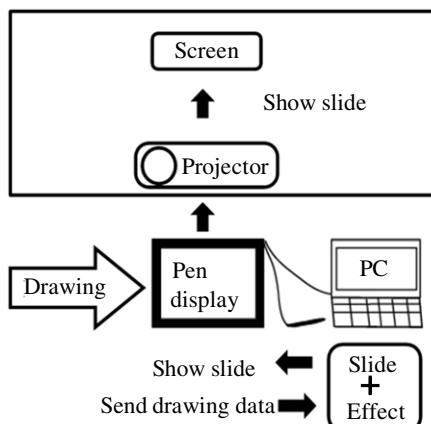


Fig. 11. System configuration

5.2 Visual Effects That Reflect Presenter's Action

During a presentation, the presenter's actions can attract an audience. Herein, on the basis of the results presented in Chapter 4, we have developed two visual effects that introduce 0.5-s motions and actions of a presenter, as shown in Fig. 12, as there is a gap of about 0.5 s between an utterance and an emphasis.

Therefore, we developed two visual effects that reflect a presenter's action. From the results presented in Chapter 4, it can be said that there is a gap of about 0.5 s between an utterance and an emphasis, and the developed visual effects reflect a 0.5-s action.

The Appearing Finger is one of these effects. When a presenter touches a pen display, the display shows a semi-transparent finger-shaped icon, the color of which gradually changes to black for 0.5 s. This effect is displayed to represent a finger appearing on a slide.

The second effect is a Zooming Finger. When the presenter touches a pen display, the display shows a large finger-shaped icon, which gradually becomes a smaller icon for 0.5 s. This effect represents a presenter's finger approaching the slide.

5.3 Visual Effects Control the Timing of Emphasizing Subjects

We have developed two additional effects by controlling the timing of an emphasized motion of a presenter, as shown in Fig. 13.

A Square Line is one of these effects. This effect occurs when the presenter draws an L-shaped line. When the presenter draws such a line, a semitransparent square composed of the start and end points of the line are shown.

The other effect is a Square Zoom. This effect also occurs when the presenter draws an L-shaped line. When this occurs, the square area composed of the start and end points of the line is zoomed 1.5 times.

These effects designate a presenter's drawing action. Thus, the presenters are conscious of their drawing, and the timing gap is filled naturally.

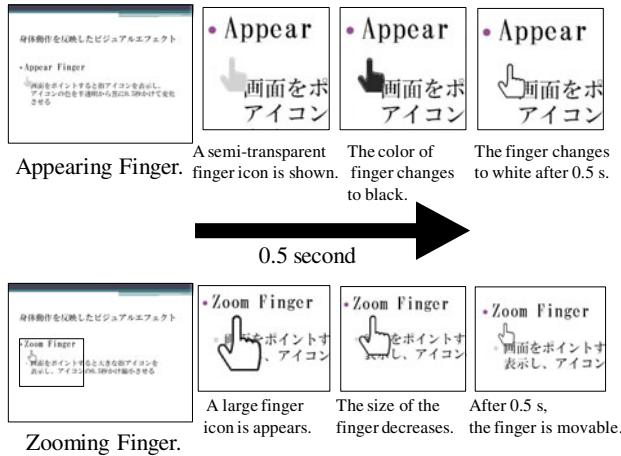


Fig. 12. Visual effects that express embodied motions

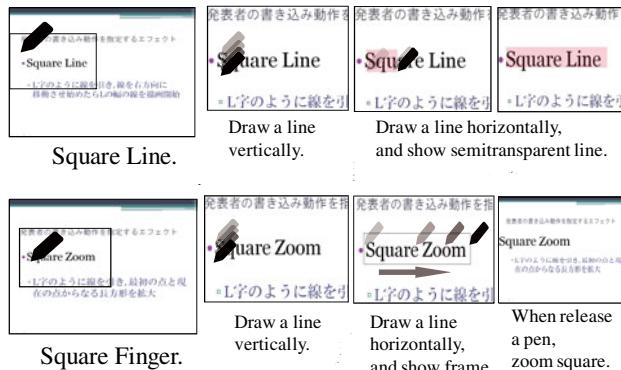


Fig. 13. Visual effects for controlling embodied timing

6 Conclusion

In this study, we developed a concept of embodying visual effects that expand the presentation motions of emphasis and indication using a pen display. First, we measured the presentation motions of pen use by focusing on the timing while in sitting and standing positions and we made it clear that an underlining motion has about a 0.5-s lag behind an utterance. Next, we evaluated the timing of underlining and explanation using an analysis by synthesis from the viewpoint of the audience, and we made it clear that the audience prefers a pattern in which the underlining motion and utterance begin at the same time. Then, on the basis of the results of our measurements and evaluation, we developed two visual effects that express embodied motions. We also control embodied timing as a prototype of the system. Concretely, the first effect expresses embodied motions (Fig. 12) and the other is the visual effect that controls embodied timing (Fig.13). Hereafter, we will confirm the effectiveness of these effects.

Acknowledgements. This work under our project “Embodied Communication Interface for Mind Connection” has been supported by “New IT Infrastructure for the Information-explosion Era” of MEXT KAKENHI. Also, our project “Generation and Control Technology of Human-entrained Embodied Media” has been supported by CREST of JST.

References

1. Tan, K.-H., Gelb, D., Samadani, R., Robinson, I., Culbertson, B., Apostolopoulos, J.: Gaze Awareness and Interaction Support in Presentations. In: ACM International Conference on Multimedia, HPL-2010-187, pp. 643–646 (2010)
2. Cao, X., Ofek, E., Vronay, D.: Evaluation of alternative presentation control techniques. In: Conference on Human Factors in Computing Systems: CHI 2005 Extended Abstracts on Human factors in Computing Systems, April 02-07, pp. 1248–1251 (2005)
3. Anderson, R., Anderson, R., Simon, B., Wolfman, S.A., VanDeGrift, T., Yasuhara, K.: Experiences with a Tablet PC Based Lecture Presentation System in Computer Science Courses. In: Proceedings of the 35th SIGCSE Technical Symposium on Computer Science Education, SIGCSE 2004, vol. 36(1) (2004)
4. Yamamoto, M., Watanabe, T.: Time Lag Effects of Line Drawing to Reading Utterance in the Emphasis and Indication of Displayed Text; Human Interface 2008, DVD-ROM, pp. 545–548 (2008) (in Japanese)
5. Watanabe, T.: Embodied Communication Technologies and Their Applications. Institute of Systems, Control and Information Engineers 49(11), 431–436 (2005)
6. Yamamoto, M., Watanabe, T.: Time Lag Effects of Utterance to Communicative Actions of Human-Robot Greeting Interaction. Journal of Human Interface Society: Human Interface in Japan 6(3), 343–350 (2004) (in Japanese)
7. K.F.: A Laser Pointer/Laser Trails Tracking System for Visual Performance; Human-Computer Interaction - INTERACT 2005 (LNCS3585). In: Costabile, M.F., Paternó, F. (eds.) INTERACT 2005. LNCS, vol. 3585, pp. 1050–1053. Springer, Heidelberg (2005)

8. Murata, Y., Shizuki, B., Tanaka, J.: Shadowgraph: A presentation tool using a pen shadow. WISS in Japan, pp. 73–78 (2008) (in Japanese)
9. Nagai, H., Yamamoto, M., Watanabe, T.: A Speech-Driven Embodied Entrainment System with Visualized and Vibratory Nodding Responses as Listener. Transactions of the Japan Society of Mechanical Engineers in Japan 75(755), 163–171 (2009) (in Japanese)
10. Wesugi, S., Kubo, T., Miwa, Y.: Tool-type interface system supporting for an expansion of body image toward a remote place - Development of virtual shadow interface system. In: Proceedings of Society of Instrument and Control Engineers, pp. 912–917 (2004)

Experimental Study on Appropriate Reality of Agents as a Multi-modal Interface for Human-Computer Interaction

Kaori Tanaka², Tatsunori Matsui², and Kazuaki Kojima¹

¹ Graduate School of Human Sciences, Waseda University

² Faculty of Human Sciences, Waseda University

2-579-15, Mikajima Tokorozawa Saitama, 359-1192, Japan

mono_season101@asagi.waseda.jp, matsui-t@waseda.jp,
koj@aoni.waseda.jp

Abstract. Although humanlike robots and computer agents are fundamentally recognized as familiar, considerable similar external representation occasionally reduces their familiarities. We experimentally investigated relationships between the similarities and the familiarities of multi-modal agents which had face and voice representation, with the results indicating that similarities of the agents didn't simply increase their familiarities. The results in our experiments implied that external representation of computer agents for communicative interactions should not be very similar to human but appropriately similar in order to gain familiarities.

Keywords: Multi-modal agent, face, voice, similarity, familiarity, uncanny valley.

1 Introduction

Computers are indispensable in a variety of contexts and tasks in the modern society. It is therefore an important and general issue in human-computer studies to support novices as end users in interacting with computers so that much more people gain benefits from computers. One of the approaches to facilitate the use of computers by novices is to make computers more familiar for human [1]. To do so, it is effective to implement a computer which has humanlike representation such as a human-shaped appearance or behavior showing a politeness. People essentially sense familiarity toward things similar to human. Thus, computers with humanity representation can be more likely to successfully interact with novices.

Although computers similar to human are recognized as familiar, considerable similar external representation occasionally reduces the familiarity. Mori [2] documented the *uncanny valley* which hypothesized that robots' familiarities perceived by human increases as the robots appear more humanlike, even though, a robot with a quite human-like appearance gives us a sense of strangeness so that its familiarity drastically decreases to the bottom of a valley. The hypothesis of the uncanny valley has not been sufficiently verified yet. However, some studies demonstrated cases where such human-like appearances of robots or computer agents

reduce their familiarities. For example, Komatsu and Yamada [3] experimentally indicated that familiarities of agents were decreased by a gap between what users had expected for the agents before interactions and actual functions of the agents. According to those facts, agents which are highly similar to human could spoil their familiarities due to their design elements such as functions. To implement familiar computer agents, we must consider two viewpoints: design of the agents and attribution for them. The former viewpoint of the design refers to how the agents are constructed and how they work, and the latter of the attribution how people recognize the agents. The uncanny valley implies that human-likeness of agents in the viewpoint of the design doesn't necessarily increase that in the attribution from human. From the aspect of familiarities, attractive computational agents may not be very similar to human, but appropriately similar to some extent. It is hence an important task in studies of computer agents to understand relationships between similarities to human and familiarities of agents.

For the goal to provide a basis to implement familiar human-like computers, current study obtained empirical data to describe relationships between similarities and familiarities of computer agents. We conducted experiments where participants were asked to interact with computer agents whose similarities of external representation were controlled, and to evaluate their familiarities. We adapted multi-modal agents as targets of the evaluations which have external representation presented through two communication modalities: facial representation in the visual modality and voice in the auditory modality. Faces and voice are essential and precedent elements in impressions of a person in human communication [4]. Thus, the multi-modal agents in our experiments can be considered as one of typical humanlike agents for communicative interactions with people. We controlled the degrees of similarities to human in representation in each modality. Our experiments were performed in two phases. In the first phase, mono-modal agents which had either of faces or voice were preliminary evaluated. The multi-modal agents were studied in the second phase. The participants in the experiments rated human-likeness, familiarity, and impression of each agent. The influences of the faces and voice on the familiarities of the agents were also examined.

In this paper, we briefly explain the first phase in our experiments, which investigated the mono-modal agents in Section 2. We then describe the experimental method and results of the second phase for the multi-modal agents, which followed by discussion of the relationships between similarities and familiarities of agents in Section 3. We finally conclude the current study in Section 4.

2 Preliminary Studies

This section briefly explains the first phase in our experiments, where mono-modal agents with faces or voice were preliminary studied. The purpose in this phase was to produce faces and voice constituting multi-modal agents in the second phase described below, and obtain their empirical data. We therefore asked participants to evaluate similarities and familiarities of agents whose external representation was composed controlling similarities to human.

2.1 Evaluation of Agents with Faces

We first empirically evaluated mono-modal agents with humanlike faces. Images of the faces covering surfaces of the agents were produced controlling similarities in features of optical and morphological validity. The optical validity denotes the solidity of a target shape, and the morphological validity positions of components on the target. Four face images shown in Fig. 1 were generated, each of which was high/low in the optical/morphological validity. In the generation of the four images, we first created an averaged face by combining face images of 62 men as the one which were high in optical and morphological validity (we refer to it as O_hM_h). The other three faces were then created according to facial feature points extracted from O_hM_h . O_hM_l was composed by putting face parts provided from Ultimate Flash Face, a web application to generate face-montages, on their respective feature points. Those face parts were solidly represented (high in optical validity) but their shapes were different from those of O_hM_h (low in morphological validity). In the same way, O_lM_h was composed with Charappeal, another face-montage software. We selected and arranged face parts whose shapes were similar to those on O_hM_h (high in morphological validity), but the parts were not solidly represented (low in optical validity). Finally, O_lM_l was created by drawing dots or a lines on the feature points of the eyes, mouth and contour, so that it was dissimilar both in optical and morphological validity.

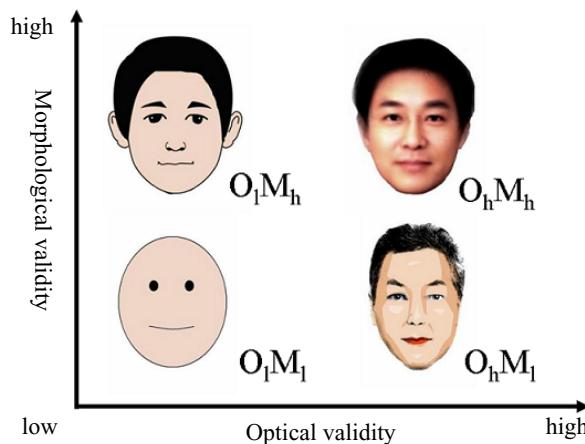


Fig. 1. Face images

We in advance asked 23 undergraduates to evaluate similarities of the four faces to human through pairwise comparison using the Visual Analog Scale. A week later, we asked the undergraduates to evaluate impressions of agents, each of which had either of three faces (targets) other than the most dissimilar face (base, which was actually O_lM_l). The agents were embedded in a web application of virtual weather report service and told weather information a user requested. The undergraduates evaluated

an impression of each target agent through comparison with the base agent with the Semantic Differential method after they had in turn interacted with the base and target agents. Evaluating items in the SD method were introduced from the method of the big five personality traits. Fig. 3 shows the similarities of the four faces, and Fig. 4 shows factors of conciliatory, activeness and familiarity extracted from the factor analysis of impressions evaluated through the SD method. As Fig. 3 indicates, O_hM_h was evaluated as most similar to human. However, O_hM_h did not gain higher scores than the others, rather rated the lowest scores among the three target agents in all of the three factors. Those results indicate that the similarities of the faces didn't necessarily increase their familiarity.

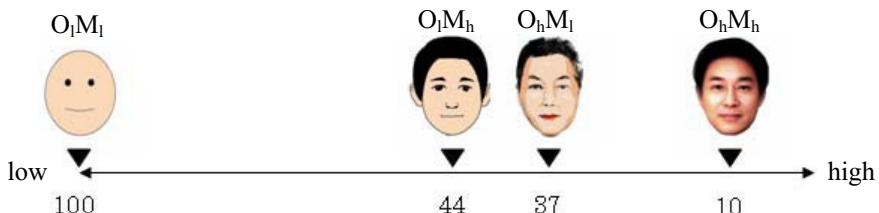


Fig. 2. Simiralities of face image

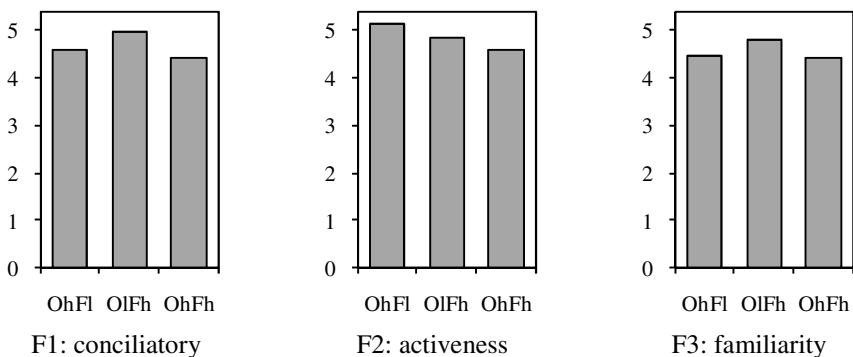


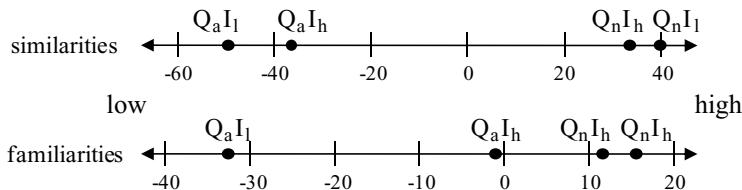
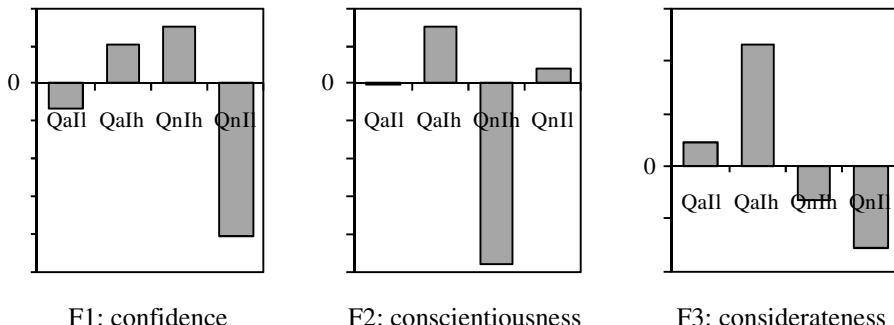
Fig. 3. Factor scores of target face agents

2.2 Evaluations of Agents with Voice

In the same way as the face agents, we evaluated mono-modal agents with voice speech. The voice was produced controlling features of the quality and intonation. The factor of voice quality had two levels: natural voice and the artificial-synthesized voice. The natural voice was produced by recording and editing speech by the experimenter, and the artificial voice was generated by Open JTALK, a voice-synthesizing software. Another factor of the voice intonation was set as high or low by strengthening or weakening that of the natural and artificial voice. Thus, we produced four types of voice as shown in Table 1.

Table 1. Voice

		Quality	
		Natural	artificial
Intonation	high	$Q_n I_h$	$Q_a I_h$
	low	$Q_n I_l$	$Q_a I_l$

**Fig. 4.** Similarities and familiarities of voice**Fig. 5.** Factor scores of voice agents

Similar to the evaluation of the face, we first asked 24 undergraduates to evaluate similarities and familiarities of the voice through the pairwise comparison. We then asked 20 of undergraduates, graduates or adults to separately evaluate impressions of each agent, which spoke by using one of the voice in the virtual weather report service, with the Semantic Differential method. These agents only made speech without any appearances. Fig. 4 shows the similarities and familiarities of the voice, and Fig. 5 shows factors from the factor analysis of the SD method. As indicated in Fig. 4, the similarities of the voice made of the natural human speech are higher than the artificial one. Although that was same in the familiarities, the order of the voice in the familiarities was not the same as that in the similarities. $Q_n I_l$ was evaluated as more similar but less familiar than $Q_n I_h$. As Fig. 5 indicates, $Q_n I_l$ also rated low scores in all of the three factors, indicating that it was evaluated as negative impressions. Those results indicate that the similarities of the voice didn't necessarily increase its familiarities.

3 Experimental Study

This section describes the second phase in our experiments to examine multi-modal agents with face and voice representation. We studied familiarities of the agents from following three viewpoints through this phase. The first viewpoint was relationships between similarities of the agents and their familiarities, the second was influences of faces and voice comprising the agents on the similarities and familiarities, and the third was relationships between impressions of the agents and their similarities or familiarities.

3.1 Method

As mentioned, we experimentally studied multi-modal agents which comprised of the faces and voice produced in the first phase. To represent the similarities of the agents' faces and voice, we replace labels for the faces and voice to those in the Table 2. The numerals added to new labels indicate the orders of the similarities in the faces or voice. A label M^x_y indicates a multi-modal agent comprised of a face F_x and voice V_y . In the experiment of this phase, we asked 37 undergraduates to evaluate impressions of each of the 16 agents with the SD method identical to the first phase after single interaction with the agent on the virtual weather report service.

Table 2. Multi-modal agents, and their comprising faces and voice

		Faces			
		F_1	F_2	F_3	F_4
Voices	V_1	M^1_1	M^2_1	M^3_1	M^4_1
	V_2	M^1_2	M^2_2	M^3_2	M^4_2
	V_3	M^1_3	M^2_3	M^3_3	M^4_3
	V_4	M^1_4	M^2_4	M^3_4	M^4_4

Because a large number of pairs had to be evaluated through the pairwise comparison adopted in the first phase, we abandoned it in this phase. Similarities and familiarities of the agents were estimated by Thurstone's Paired Comparison based on evaluating values of some adjectives from the SD method. The similarities were estimated from evaluating values of "warm" and "having respondible", and the familiarities were from "attractive" and "friendly". The first and second viewpoints mentioned above were described according to the similarities and familiarities newly estimated. Furthermore, we performed the factor analysis of the impressions evaluated through the SD method in order to investigate relation between the personality factors and the similarities or familiarities

3.2 Results

Fig. 6 shows the new similarities and familiarities of the multi-modal agents. The figure obviously indicates that the similarities and familiarities didn't simply

correspond but differed from each other. Another interesting fact in the figure is that two groups of higher or lower agents in the similarities were the same as those in the familiarities. Every agent in the higher group had the first or second highest voice in the similarities. This may indicate that the similarities of the voice influenced on both the similarities and familiarities of the multi-modal agents in some ways.

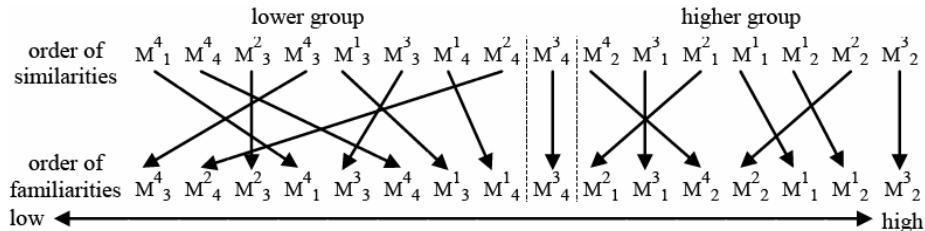


Fig. 6. Orders of multi-modal agents on similarities or familiarities

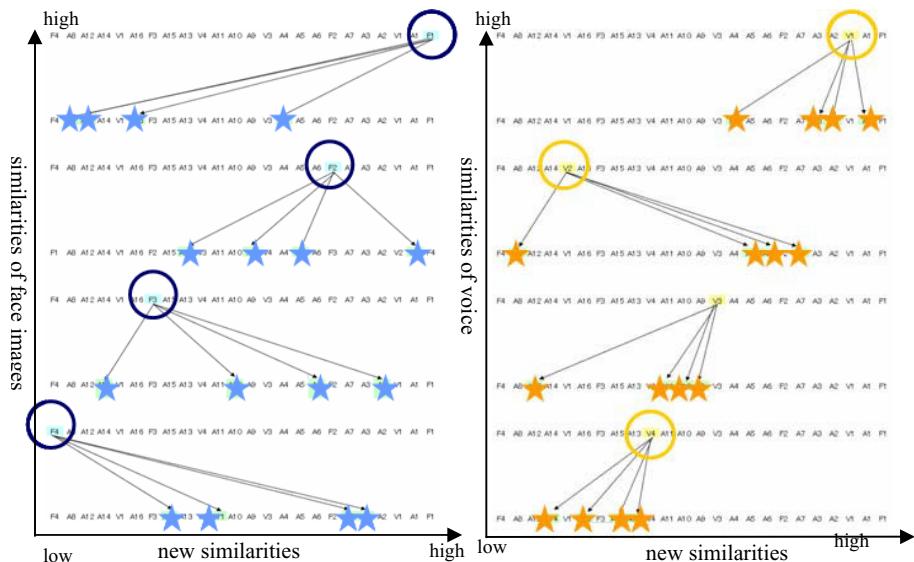


Fig. 7. Relationships between multi-modal agents and their faces on similarities

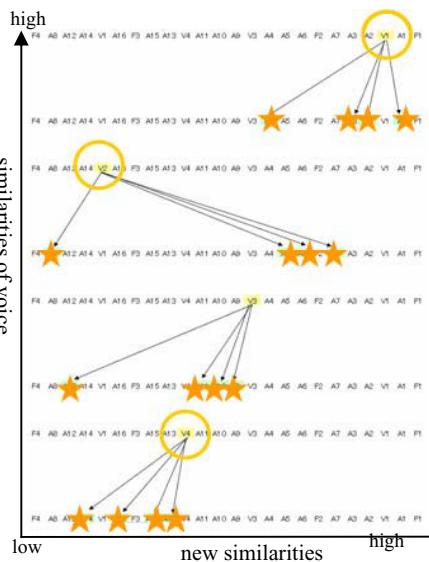


Fig. 8. Relationships between multi-modal agents and their vice on similarities

Fig 5, 6, 7 and 8 show relationships between the mono- and multi-modal agents in the similarities or familiarities. In these figures, vertical axes represent the similarities of the face or voice measured in the first phase. Horizontal axes represent the new similarities in Fig 5 and 6, and the new familiarities in Fig 7 and 8. Each line in the figures includes the agents arranged according to orders of the new similarities or familiarities. Each circle in a line marks a mono-modal agent with a face or voice, and each set of four stars linked with the circle marks multi-modal agents which include the face or voice. In most cases, the similarities and familiarities of the multi-modal

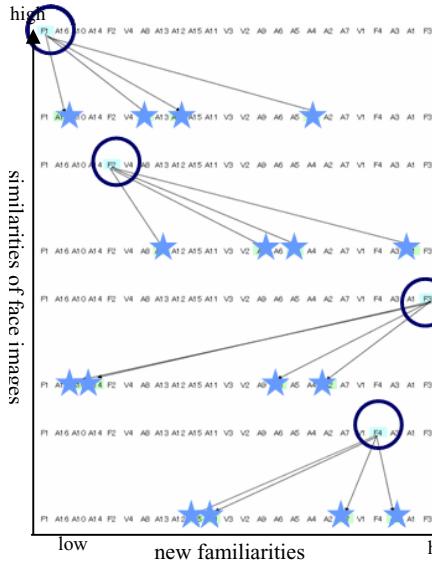


Fig. 9. Relationships between multi-modal agents and their face on familiarities

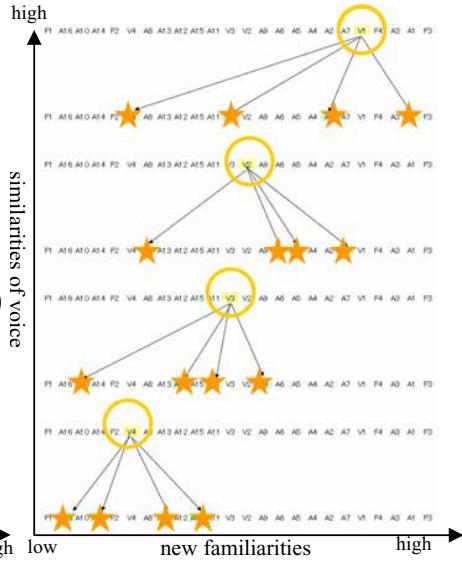


Fig. 10. Relationships between multi-modal agents and their vice on familiarities

agents represented in the horizontal axes were not consistent with the similarities of face or voice agents in the vertical axes, which fact again indicates that the similarities of the faces or voice didn't necessarily increase the similarities and familiarities of the multi-modal agents. Thus, we had to further investigate the relationships between the similarities and familiarities. Fig. 8, 9 and 10 present scatter diagrams of the similarities and familiarities. Spearman's rank correlation coefficients between the similarities and familiarities revealed that they had a positive correlation in the voice and multi-modal agents ($r=0.92$), and a negative correlation in the faces agents ($r=-0.90$).

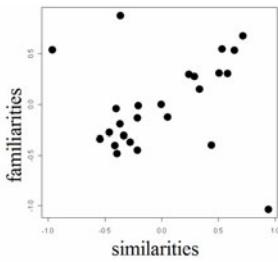


Fig. 11. Scatter diagram of all agents

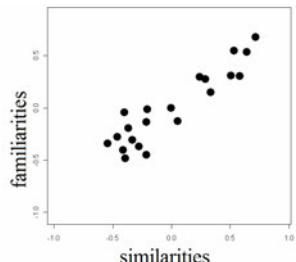


Fig. 12. Scatter diagram of multi-modal and voice agents

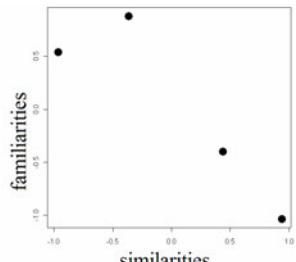


Fig. 13. Scatter diagram of face agents

Table 3. Factor loadings of each evaluating items

	Factor1	Factor2	Communality
	cooperation	leadership	
extrovert	0.161	0.950	0.929
attractive	0.944	0.258	0.958
cautious	0.844	-0.452	0.916
confident	-0.036	0.985	0.972
companionable	0.893	0.252	0.861
cooperative	0.955	0.188	0.948
conscientious	0.907	-0.370	0.959
considerate	0.859	-0.498	0.986
strong	-0.522	0.820	0.945
having respondible	0.933	-0.213	0.915
Soft	0.872	-0.263	0.829
friendly	0.790	-0.195	0.662
Successful	0.898	0.296	0.893
diligent	0.936	-0.264	0.945
straight	0.912	-0.371	0.970
Proportions of variance	66.269	24.982	
Cumulative proportions	66.269	91.251	

Two of “cooperation” and “leadership” factors were extracted (the cumulative proportion was 91.25 %) from the factor analysis of the SD method. Table 3 indicates the factor loadings for the two factors. However, we found no clear relationships between those factors and the similarities or familiarities. It is one of our future work to study computer agents from the viewpoint of impressions.

3.3 Discussion

The results described above illustrated that the similarities to human of the multi-modal agents didn't merely increase their familiarities. Furthermore, the similarities constituting the multi-modal agents of the voice were correlated to their similarities and familiarities, although the faces which were also constituted the agents were not.

Why didn't the faces similar to human, which represented the rich humanlike surfaces, necessarily increase the familiarity of its agents? That may be due to limitations of human information processing in interpersonal cognition. In formulating impressions of a person, people don't process entire information of the person but constrain the information to be processed according to social or cognitive factors, such as interpersonal relationships or purposes in interactions. Such constraining can occur when interacting with computer agents. The situation was to listen to weather information in our experiment, so that the participants may have insufficiently focused on the face information of the multi-modal agents.

4 Conclusion

To provide a basis for designing appropriate external representation of humanlike computers from the aspect of familiarities perceived by human, this study investigated

relationships between the similarities and familiarities of multi-modal agents with faces and voice as external representation. We conducted experiments to evaluate the similarities, familiarities and impressions of the faces, voice and agents comprised of them. The results indicated that the similarities of the agents didn't simply increase their familiarities. Thus, we believe that it proved to be a good idea to design computers whose external representation is appropriately similar to human as long as they are used in communicative interactions and required to gain familiarities. One important task in the future works is to explore human factors in the recognition of the familiarities of computer agents.

References

1. Reeves, B., Nass, C.: *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Center for the Study of Language and Information, California (1996)
2. Mori, M.: *The Uncanny Valley*. K. F. MacDorman & T. Minato, Trans., Energy, vol. 7, pp. 33–35 (1970)
3. Komatsu, T., Yamada, S.: Adaptation Gap Hypothesis: How Differences between Users' Expected and Perceived Agent Functions Affect their Subjective Impression. *Journal of Systemics, Cybernetics and Informatics* 9, 67–74 (2011)
4. Kawanishi, C.: *The Integration Process of Interpersonal Information on Impression Formation*. Kazamashobo Co., Ltd., Tokyo (2002) (in Japanese)

Author Index

- Abdullah, Natrah I-17
Abe, Kiyohiko II-176
Abe, Shuya III-240
Abou Khaled, Omar II-185, II-222, III-165
Abraham, Chon IV-141
Abril-Jiménez, Patricia III-129
Abusham, Eimad E.A. II-169
Adams, Candice IV-3
Adikari, Sisira I-25
Ahn, Hyeong-Joon III-386
Ahn, Jong-gil II-304
Ai, Renlong II-387
Aihara, Kenro III-271
Akahane, Katsuhito II-3
Akaho, Kengo III-357
Alcañiz, Mariano I-423
Alnizami, Hanan II-533
Altaboli, Ahamed I-35
Alvarez, Haydee III-129
Amraji, Nitish I-539
Andrade, Vladimir I-54
Andreoni, Giuseppe I-406
Andrich, Rico II-396
Angelini, Leonardo II-185
Antonya, Csaba II-204
Aoki, Ryosuke II-194
Aoyagi, Saizo III-465
Aoyama, Hifumi III-84
Aoyama, Hisae I-341
Arafat, Sachi III-367
Arredondo Waldmeyer, María Teresa III-129, IV-219
Asan, Onur II-13
Asano, David II-133
Ash, Jordan III-137
Auernheimer, Brent IV-394
Babes, Monica III-137
Bachman, Mark I-107, IV-151
Balasubramanian, Venkatesh II-446
Baños, Rosa I-423
Baragaño Galán, José R. III-129
Baranauskas, M. Cecília C. I-72
Barbuceanu, Florin II-204
Barnes, Julie I-432
Barrett, Jonathan IV-580
Barroso, João III-293
Bartoschek, Thomas II-71
Bashir, Housam K. II-169
Beheshti, Jamshid IV-541, IV-590
Bengler, Klaus II-23, II-125, III-376
Benoit, David Michel II-537
Bernstein, Stefan I-618
Beznosyk, Anastasiia IV-668
Bilal, Dania IV-549
Birinci, Murat I-547
Biswas, Gautam IV-580
Biswas, Pradipta IV-425
Bizik, Kai I-165
Black, Alan II-358
Blackwell, Nicole II-547
Blattner, Andreas II-125
Blažica, Bojan III-519
Block, Micha III-474
Blumberg, Malte III-38
Böck, Ronald III-603
Boegel, Bastian II-537
Bomsdorf, Birgit I-155
Bonarini, Andrea IV-649
Botella, Cristina I-423
Boulabiar, Mohamed-Ikbel II-214
Bowden, Chris IV-443
Braun, Andreas III-147, III-205
Breiner, Kai I-165, I-299
Breyer, Matthias IV-520
Bricault, Ivan IV-202
Britton, Douglas IV-167
Brooks, Laurence IV-456
Bubb, Heiner III-376
Bucki, Marek IV-202
Buckwalter, J. Galen II-474
Burger, Thomas II-214
Burkhardt, Dirk IV-520
Byer, David II-325
Camez, Edgar III-411
Campbell, John I-25

- Campos, Miguel III-155
 Campos, Pedro III-155
 Caon, Maurizio II-185
 Carrino, Francesco II-222
 Carrino, Stefano II-185, III-165
 Casillas, Raúl III-3
 Castilla, Diana I-423
 Cearreta, Idoia III-525
 Cerrato, Marcus Vinicius I-471
 Chaim, Marcos Lordello I-471
 Chalon, René II-500
 Chamaret, Damien II-158
 Chan, Foong-Yeen Donny III-329
 Chance, Eric II-474
 Chang, Huang-Ming III-175, III-559
 Chang, Liang I-555, I-670
 Chang, Teng-Wen I-82, II-53, II-409, III-250
 Chao, Dingding III-280
 Charfi, Selem I-175
 Charassis, Vassilis III-367
 Chattaraman, Veena II-533
 Chen, Chun-Lien IV-570
 Chen, Jai-Jung I-82
 Chen, Jie I-677
 Chen, Shang-Liang IV-570
 Chen, Shao-Nung IV-86
 Chen, Sheng-Han II-409
 Chen, Sherry Y. IV-27
 Chen, Ting-Han III-185
 Chen, Wei-Chia III-21
 Chen, Yihsiu III-13
 Chen, Yun-Yao IV-570
 Cheng, Yi-Ju IV-131
 Cheng, Yun-Maw III-21
 Chiang, Chen-Wei I-45, III-30
 Chiu, Kuo Chung II-491
 Cho, Hyunchul II-119
 Cho, Young Suk IV-192
 Choe, Jaeho III-401
 Choi, Ji-Hye I-637, I-653
 Choi, Jongwoo III-452
 Choi, Seung-Hwan II-231
 Chou, Chin-Mei I-397
 Chu, Chi Nung III-288, IV-35, IV-435
 Chu, Gene III-288
 Chu, Keng-Yu II-248
 Chu, Shaowei II-238
 Chuang, Chih-Fei IV-439
 Cobb, Sue III-483
 Cockton, Gilbert III-490
 Coelho, Paulo III-293
 Cohen, Gal III-137
 Coninx, Karin IV-668
 Cooper, Eric W. III-535
 Coppin, Gilles II-214
 Cortés, José Luis IV-255
 Costa, Fiammetta IV-649
 Costa, Paulo III-293
 Couvreur, Christophe III-418
 Cremers, Anita II-575
 Dadgari, Darius II-331
 Damböck, Daniel III-376
 Dandekar, Heramb III-411
 Datta, Sambit I-185
 David, Bertrand II-500, III-301
 D'Cruz, Mirabelle III-483
 de Barriosuevo, Arturo Díaz IV-219
 de Beer, Lorraine III-500
 Dees, Walter III-195
 de Moraes Rodrigues, Roberto Leite I-471
 Deng, Qingqiong I-670
 Deng, Xiaoming I-555
 Depradine, Colin II-325
 Diebold, Philipp I-165
 Dittmar, Anke I-194
 Djamasbi, Soussan I-331, IV-245
 Do, Ellen Yi-Luen II-547, IV-192
 Donker, Hilko III-38
 Druin, Allison IV-559
 Drury, Colin G. I-397, II-275
 Duan, Fuqing I-660, I-670
 Duguleana, Mihai II-204
 Edlinger, Günter II-417
 Edlin-White, Rob III-483
 Ei, Hitomi IV-413
 Ekman, Inger I-351
 Ellis, Rebekah IV-549
 EL-Qirem, Fuad III-490
 Engel, Jürgen I-204
 Esteves, Marcela I-54
 Everard, Andrea IV-58, IV-235
 Ezzedine, Houcine I-175
 Fahn, Chin-Shyurng II-248
 Fails, Jerry Alan IV-559
 Fan, Kuo-Kuang III-259

- Fang, Xiaowen IV-659
 Feki, Mohamed Ali I-229
 Fennell, Antoinette IV-443
 Fernandes, Hugo III-293
 Fernandez-Llatas, Carlos I-214
 Fernández, Cristóbal IV-255
 Ferran, Nuria III-510
 Feuerstack, Sebastian I-221
 Fiddian, Thomas IV-443
 Fischer, Holger II-100
 Floyde, Anne III-483
 Folmer, Jens IV-297
 Forbrig, Peter I-194, I-204, I-248, I-319, IV-510
 Forutanpour, Babak III-48
 Fox, Michael A. IV-307
 Friedman, Nizan IV-151
 Friesen, Rafael I-121, III-593
 Frommer, Jörg III-593
 Fuji, Masahiko IV-21
 Furukawa, Hiroyuki IV-159
 Furukawa, Masashi IV-354
 Furukawa, Yutaroh I-588
 Furuta, Kazuo III-280
 Furuyama, Nobuhiro II-259
- Gabbouj, Moncef I-547
 Gacimartín, Carlos I-443
 Gagliardi, Chiara I-406
 Gandy, Maribeth IV-167
 Gao, Fei III-59
 Gao, Qin I-450, III-59
 Gao, Song-feng IV-688
 Garay-Vitoria, Nestor III-525
 Garbarino, Maurizio IV-649
 Garreta-Domingo, Muriel I-136, III-510
 George, Sébastien IV-12
 Germanakos, Panagiotis I-388
 Ghinea, Gheorghita II-427
 Gilbert, Juan II-533
 Gill, Sukhpreet IV-394
 Glodek, Michael III-603
 Godon, Marc I-229
 Gonçalves, Júnia I-63
 Gong, Yang IV-182
 Gower, Michael I-248
 Grau, Stefan I-155
 Grønli, Tor-Morten II-427
 Grundlehner, Bernard II-481
 Guger, Christoph II-417
- Guha, Mona Leigh IV-559
 Guldogan, Esin I-547
 Guo, Ping I-562
- Haase, Matthias III-593
 Hadjileontiadis, Leontios III-293
 Hamada, Takeo II-3
 Hamberger, Werner II-125
 Han, Ji-Hyeong II-231
 Han, Tack-don I-637, I-653
 Han, Yanjun I-555
 Hanada, Ryoko II-259
 Hansen, Jarle II-427
 Harbert, Simeon IV-167
 Harley, Linda IV-167
 Harrelson, Heather III-411
 Hashimoto, Yasuhiro IV-328
 Haslbeck, Andreas II-23
 Hatano, Akira I-628
 Hayakawa, Eiichi IV-21
 Hayashi, Yugo II-513, IV-384
 He, Ning I-571
 Heidemann, Gunther I-618
 Heikkinen, Kari I-359
 Hermann, Thomas II-110
 Hernández, José Alberto I-443
 Hess, Steffen I-238
 Hijikata, Yoshinori IV-318
 Hippler, Rachelle Kristof II-33
 Hiramatsu, Mizuko III-84
 Hobbs, Michael I-185
 Höger, Rainer III-577
 Holzner, Clemens II-417
 Honda, Daisuke III-230
 Hong, Chao-Fu IV-77
 Hong, Kwang-Seok II-464, III-569
 Hope, Tom IV-328
 Horii, Satoshi III-84
 Hornung, Heiko I-72
 Howley, Iris II-341
 Hrabal, David III-603
 Hsiao, Chih-Pin II-43
 Hsu, Chia-Ling IV-77
 Hsu, Chih-Hsiang IV-453
 Hsu, Yung-Chi IV-27
 Hu, Xiaoyan I-580
 Huang, Ho-Chuan IV-570
 Huang, Hsiu-Shuang IV-570
 Huang, Nancy IV-439
 Huang, Yu-Chun Annester I-82

- Huang, Zhao IV-456
 Hudasch, Martin I-155
 Hung, Tien-Hsin I-82
 Huseyinov, Ilham N. IV-39
 Hwang, Bong-Ha III-391
 Hwang, Sheue-Ling III-435, IV-439
 Hwang, Wonil III-386
 Hwang, Yu Ting IV-35
 Hwangbo, Hwan III-391
 Hymel, Alicia M. IV-580
 Hyrkkänen, Ursula III-69
- Ichinomiya, Takako II-259
 Ichiyanagi, Yuki III-535
 Igata, Nobuyuki III-240
 Ihara, Masayuki II-194
 Ike, Shinichi II-436
 Imtiaz Ahmad, Muneeb III-543
 Ingold, Rolf II-222, III-165
 Inoue, Masashi II-259
 Inoue, Satoru I-341
 Irino, Toshio II-259, IV-607
 Ishida, Hirotugu I-588, I-628, IV-174, IV-210
 Ishida, Toru II-565
 Ishidou, Takashi IV-413
 Ishii, Hirotake III-465, IV-403
 Ishii, Naohiro I-608
 Islam, Md. Zahidul II-294
 Ito, Kyoko IV-48
 Ito, Yuki IV-48
 Itoh, Kazunori IV-97
 Izuhara, Ritsuko IV-237
 Izumiya, Akira IV-226
- Jacob, Robert J.K. II-593
 Jaffee, Samuel D. I-432, II-33
 Jalal, Sameen III-137
 Jang, Bong-gyu III-76
 Jang, Seung-Cheol IV-360
 Javahery, Homa I-248
 Jayaraman, Srinivasan II-446
 Jeon, Myounghoon II-523
 Ji, Yong Gu III-391
 Jiang, Heng II-53
 Jin, Beom Suk III-391
 Jin, Byungki IV-257
 Jin, Hong Zhe IV-403
 Jin, Zhaoxia Janet III-329
 Jog, Jayraj II-148
- Johnson, Brian R. II-43
 Johnson, Nathan IV-336
 Johnson, Steve II-62, IV-68
 Jokisch, Markus II-71
 Jones, Brian M. IV-58, IV-235
 Joo, Jaekoo I-368
 Jorge, Joaquim III-155
 Jung, Eui S. III-401, III-426
 Jung, Hee-Seok III-386
 Jung, Helge II-100
 Junglas, Iris IV-141
- Kaelber, Claus I-258
 Kagami, Shingo II-194
 Kallenbach, Jan IV-466
 Kallinen, Kari I-351, IV-466
 Kalwar, Santosh Kumar I-359
 Kamieth, Felix III-147, III-205
 Kaneko, Ryuhei IV-346
 Kang, Sunghyun R. IV-285
 Kanno, Taro III-280
 Karatsu, Yutaka II-194
 Karpov, Alexey II-454
 Karthik, K. III-552
 Kato, Hirokazu III-357
 Kawahara, Hideki IV-607
 Kawai, Katsuya III-357
 Kawakami, Takashi IV-617
 Kawano, Show IV-403
 Kayama, Mizue IV-97
 Keller, Christine I-288
 Khairat, Saif IV-182
 Kienle, Martin III-376
 Kim, Dong-Kyu II-464
 Kim, Gerard J. II-304, III-76
 Kim, Hee-Cheol IV-476
 Kim, Ho-Won IV-476
 Kim, Hyungsin IV-192
 Kim, Jaewhan IV-360
 Kim, Ji Yeon III-391
 Kim, Jong-Hwan II-231
 Kim, Jung Hyup I-378
 Kim, Kyong-ho III-452
 Kim, Laehyun II-119
 Kim, Namhun I-368
 Kim, Nam-Hyo III-386
 Kim, Sang-Hwan III-411
 Kim, Seong M. III-401
 Kim, Taeil III-426
 Kimioka, Gimpei II-81

- Kimura, Hiroshi IV-354
 Kimura, Masaomi I-588, I-628, IV-174, IV-210, IV-226
 Kipyatkova, Irina II-454
 Kirisci, Pierre T. IV-443
 Kita, Yusuke III-213
 Klein, Brandi A. II-33
 Kleindienst, Jan III-418
 Klenner-Moore, Jayne IV-61
 Klompmaker, Florian II-90, II-100
 Klopfer, Dale S. I-432, II-33
 Kobayashi, Minoru II-194
 Koga, Hiroyuki I-461
 Kogure, Yuri IV-413
 Kojima, Kazuaki II-613, IV-384
 Kojo, Akira I-588
 Kolski, Christophe I-175, III-301
 Komischke, Tobias I-92
 Konuma, Mizuki IV-413
 Koshi, Shoko II-314
 Koshiba, Hitoshi III-271
 Krahmer, Emiel II-575, III-543
 Krause, Michael II-23
 Kryssanov, Victor V. II-513, III-535
 Kuijper, Arjan I-480, IV-520
 Kuramoto, Itaru II-351
 Kurihara, Kazutaka II-557
 Kurihara, Lisa IV-607
 Kweon, In-So II-294
 Kwon, Hyeong-Joon III-452
 Kwon, Hyeong-Oh III-569
 Kwon, Wi-Suk II-533
 Labský, Martin III-418
 Lang, Helmut II-537
 Langdon, Patrick IV-425
 Lange, Christian II-125
 Lange, Julia III-593
 Large, Andrew IV-541, IV-590
 Larrabeiti, David I-443
 Lawo, Michael IV-443
 Le Bellego, Gaël IV-202
 Lee, Chao-Lung III-21
 Lee, Chil-Woo II-268, II-294
 Lee, Hyunglae IV-484
 Lee, Jun-Sung II-268, II-294
 Lee, Myeong-Hee IV-328
 Lee, Myonghee III-76
 Lee, Seung Jun IV-360
 Lee, Ying-Lien IV-453
 Lekkas, Zacharias I-388
 Leou, Yea-Mei IV-570
 Leventhal, Laura Marie I-432, II-33
 Levin, Daniel T. IV-580
 Lewthwaite, Sarah III-483
 Li, Yibai IV-336
 Li, Yueqing II-62, IV-68
 Li, Zhanhuai III-220
 Liang, Chao I-660
 Liang, Rung-Huei III-175, III-559
 Liang, Sheau-Farn Max I-98
 Lichtenberg, Sam III-137
 Lim, Youngjae III-426
 Limbrecht, Kerstin III-603
 Lin, Brian Tsang-Wei III-435
 Lin, Chien-Tsen I-98
 Lin, Hsin-Hou III-185
 Lin, Jui-Feng I-397, II-275
 Lin, Mu-Hua IV-77
 Lin, Sheng Ta II-491
 Lin, Ya-Li III-311
 Lin, Yi-Chu III-559
 Lin, Yingzi I-35
 Lin, Yi-Quan I-397
 Lin, Yu-De I-397
 Lin, Yuh-Chang IV-77
 Link, Jasmin III-474
 Littman, Michael III-137
 Liu, Cha-Lin II-53
 Liu, Cheng-Li IV-490
 Liu, Deyun IV-476
 Liu, Jane W.-S. III-107
 Liu, Jie IV-368
 Liu, Xianging II-140
 Liu, Yen Ping II-491
 Loiacono, Eleanor IV-245
 López, Marta I-136
 Loveless, Steve IV-394
 Lu, Cindy IV-375
 Lu, Ke I-571
 Lu, Szu-Chia II-547
 Lu, Yin III-220
 Luh, Ding-Bang IV-86
 Luo, Ruiyi I-598
 Luo, Wei Wen II-491
 Luyten, Kris IV-668
 Ma, Jin I-562
 Macek, Tomáš III-418
 Maciejewski, Anthony A. I-450

- Mackenzie, Lewis III-367
 Maeda, Atsuhiko II-194
 Maehgashi, Akihiro IV-384
 Maekawa, Yasuko III-84
 Magennis, Mark IV-443
 Maier, Andreas I-238
 Majima, Yukie III-84
 Mallem, Malik II-140
 Mangas, Juan-Antonio III-510
 Marivate, Vukosi III-137
 Märtin, Christian I-204, I-258
 Maruyama, Minoru II-133
 Massaki, Hiroyasu II-259
 Matos, Gilberto I-142
 Matsui, Tatsunori II-613
 Matsunobe, Takuo II-378
 Matsuuchi, Naohisa II-436
 Matteucci, Matteo IV-649
 Mattheij, Ruud III-500
 Mayer, Jürgen II-23
 Mazalek, Ali II-148
 McCoy, Scott IV-58, IV-235, IV-255
 McDonald, Craig I-25
 Md Noor, Nor Laila I-17
 Meixner, Gerrit I-165, I-299
 Mendenhall, Sam II-148
 Merlo, Mark I-107
 Metze, Florian II-358
 Miao, Kejian III-220
 Milanova, Mariofanna I-539
 Milde, Jan-Torsten I-155
 Min, Jin-Hong III-569
 Minakuchi, Mitsuru II-351
 Minker, Wolfgang I-527, II-537
 Mitchell, Stephen J. IV-394
 Mitsuishi, Takashi IV-500
 Miwa, Kazuhisa II-513, IV-384
 Miyabe, Mai II-368
 Miyagi, Kazune IV-403
 Miyamori, Shoko IV-607
 Miyao, Hidetoshi II-133
 Miyazaki, Tsuyoshi I-608
 Miyoshi, Yasuo IV-93
 Mizukawa, Makoto III-586
 Mladenić, Dunja III-519
 Mochizuki, Toshio II-557
 Mocholí, Juan Bautista I-214
 Moehrmann, Julia I-618
 Mohamad, Yehya IV-443
 Mohd. Ali, Borhanuddin II-285
 Montague, Enid II-13
 Moon, Young Joo III-391
 Mor, Enric III-510
 Moragrega, Inés I-423
 Morán, Alberto L. III-3
 Morandini, Marcelo I-471
 Morie, Jacquelyn Ford II-474
 Morita, Junya IV-384
 Mort, Greg III-321
 Mortimer, Bruce III-321
 Mosch, Christian II-537
 Mourlas, Constantinos I-388
 Moussa, Faouzi I-175, III-92
 Mu, Lin I-539
 Mugellini, Elena II-185, II-222, III-165
 Müller, Tobias II-90
 Muranaka, Noriaki IV-121
 Muraoka, Hiroki IV-226
 Murata, Kazuyoshi III-101
 Murata, Yuichi II-557
 Nabeta, Keita I-588, I-628, IV-174, IV-210
 Nagai, Takashi IV-97
 Nagata, Kazunobu III-101
 Naka, Toshiya II-565
 Nakada, Toru II-481
 Nakagawa, Takashi III-357
 Nakagawa, Yuuki IV-21
 Nakajima, Yukari III-84
 Nakamura, Yuki III-444
 Nakamura, Yumiko III-84
 Nakashima, Toyoshiro I-608
 Nakata, Keiichi I-341
 Nakatani, Yoshio III-213, III-444, IV-346
 Nam, Chang Soo II-62, IV-68
 Naranjo, Juan Carlos I-214
 Nauts, Pim II-575
 Nazemi, Kawa I-480, IV-520
 Nebe, Karsten I-114, II-90, II-100
 Nenonen, Suvi III-69
 Nessel, Valerie IV-599
 Ng, Chee Kyun II-285
 Ng, Wei Lun II-285
 Nishida, Shogo III-230, III-357, IV-48, IV-318
 Nishino, Yosuke IV-21
 Nisimura, Ryuichi IV-607
 Noordin, Nor Kamariah II-285

- Ochiai, Hideharu IV-21
 O' Connor, Joshue IV-443
 Octavia, Johanna Renny IV-668
 Oehl, Michael III-577
 Ogawa, Hitoshi II-513, III-535
 Oh, Chi-Min II-268, II-294
 Ohi, Shoichi II-176
 Ohkawa, Yuichi IV-500
 Ohkura, Michiko I-588, I-628, IV-174, IV-210, IV-226, IV-413
 Ohyama, Minoru II-176
 Oka, Mizuki IV-328
 Okamoto, Kentaro II-585
 Okamura, Tomoaki III-465
 Okazaki, Tetsuo IV-617
 Okuya, Ryo IV-210
 Oliveira, Kathia III-301
 Olivier, Hannes IV-678
 Omernick, Mark II-13
 Oobayashi, Takaaki IV-93
 Otmane, Samir II-140
 Ottaviano, Manuel IV-219
 Otto, Mirko I-121, III-593
 Ozaki, Shun II-378
- Paelke, Volker I-114
 Páez, José Manuel IV-219
 Pantförder, Dorothea IV-297
 Papanastasiou, Stylianos III-367
 Park, Gie-seo II-304
 Park, Hyesun III-452
 Park, Jaekyu III-401
 Park, James I-637
 Park, Ji-Hyung IV-484
 Park, Sehyung II-119
 Park, Sungjoon III-426
 Park, Wanjoo II-119
 Patki, Shrishail II-481
 Pavlov, Oleg IV-245
 Pecot, Katrina II-23
 Peer, Dain II-13
 Peinado, Ignacio IV-219
 Peissner, Matthias I-268, I-498
 Penders, Julien II-481
 Penstein Rosé, Carolyn II-341
 Perego, Paolo I-406
 Pestana, João III-155
 Pfister, Hans-Rüdiger III-577
 Pinkwart, Niels IV-678
 Pitakrat, Teerat I-527
- Pizzolato, Ednaldo I-221
 Plocher, Thomas III-346
 Plocher, Tom III-329
 Poirier, Franck II-214
 Polzehl, Tim II-358
 Ponnusamy, R. III-552, IV-274
 Poor, G. Michael I-432, II-33, II-593
 Popova, Severina II-23
 Porras, Jari I-359
- Qin, Hua IV-688
 Qin, Yongqiang I-507
 Quast, Holger III-418
 Quax, Peter IV-668
 Quiza, Phillip III-137
- Rakiman, Kartini III-500
 Rasooli, Amin IV-510
 Rattanyu, Kanlaya III-586
 Rau, Pei-Luen Patrick I-450, IV-688
 Rauch, Thilo I-165
 Ravaja, Niklas I-351, IV-466
 Rayan, Infantdani A. II-523
 Rebel, Matthias II-387
 Reinkensmeyer, David IV-151
 Ren, Jianfeng III-48
 Rey, Beatriz I-423
 Rhiu, Ilsun IV-257
 Riahi, Ines III-92
 Riahi, Meriem III-92
 Richard, Paul II-140, II-158
 Richardson, Kevin H. I-131
 Riedel, Johann III-483
 Riedenklau, Eckard II-110
 Ritter, Helge II-110
 Robert, Jean-Marc III-301
 Robertson, Scott IV-167
 Rodeiro Iglesias, Javier I-278, I-309
 Roelands, Marc I-229
 Roh, Yong-Wan II-464
 Romero, Maximiliano IV-649
 Ronzhin, Andrey II-454
 Rösner, Dietmar I-121, II-396, III-593
 Rothrock, Ling I-368, I-378, I-414
 Rumeau, Pierre IV-530
 Rusak, Zoltan II-204
 Ryu, Taebeum IV-257
- Saeed, Mehreen III-543
 Saka, Emmanuel IV-285

- Sakai, Akiko IV-413
 Sakata, Nobuchika III-230
 Sala, Pilar I-214
 Salim, Siti Salwah IV-627
 Samaras, George I-388
 Sandnes, Frode Eika I-643, III-21
 Santos, Caroline I-63
 Sarker, Saonee IV-336
 Sato, Makoto II-3
 Saylor, Megan IV-580
 Schels, Martin III-603
 Scherer, Stefan III-603
 Schlegel, Thomas I-288, I-618
 Schlehuber, Christian III-205
 Schmeier, Sven II-387
 Schmidt, Miriam III-603
 Schuller, Andreas I-268
 Schwenker, Friedhelm III-603
 Schwering, Angela II-71
 Seals, Cheryl IV-3
 Secore Levis, Melissa IV-266
 Seissler, Marc I-165, I-299
 Seo, Jonghoon I-637, I-653
 Serna, Audrey IV-12
 Shahid, Suleman III-500, III-543
 Shi, Xiaoming IV-694
 Shi, Yuanchun I-507, III-117, IV-368
 Shiba, Haruya II-436
 Shibuya, Yu III-101
 Shigeno, Aguri II-378
 Shih, Chi-Sheng III-107
 Shih, Sheng-Cheng II-409
 Shih, Ya-Chun IV-131
 Shim, Soo In II-533
 Shimamura, Kazunori II-436
 Shimoda, Hiroshi III-465, IV-403
 Shin, Hyunjin IV-484
 Shin, Seungjae II-119
 Shirai, Akira I-490
 Shizuki, Buntarou II-81, II-557
 Shneiderman, Ben I-3
 Shovan, Curtis III-321
 Siebert, Felix W. III-577
 Siegel, Howard Jay I-450
 Siegel, Marisa I-331
 Simon, Lajos I-498
 Sinnig, Daniel I-248
 Son, Young-Jun I-368
 Spath, Dieter I-268
 Spies, Roland II-125
 Stab, Christian I-480, IV-520
 Stuerzlinger, Wolfgang II-331
 Sun, Bo I-580
 Sun, Xianghong III-346
 Suo, Yue I-507
 Suzuki, Akihiro IV-617
 Suzuki, Shinya IV-237
 Szilvasi, Lindsy III-500
 Taghiyareh, Fattaneh IV-510
 Takacs, Barnabas I-498
 Takahashi, Tetsuro III-240
 Takahashi, Yuzo II-314
 Takao, Yuya II-603
 Takeda, Hideaki III-271
 Takeda, Hiroshi II-133
 Tanabe, Naohisa IV-113
 Tanaka, Jiro II-81, II-238, II-557
 Tanaka, Kaori II-613
 Tanaka, Sayaka IV-413
 Tanaka, Yuya IV-318
 Tang, Fengchun IV-336
 Tang, Sheng Kai II-491
 Taniguchi, Tadahiro I-461, I-490
 Tariq, Hassan III-543
 Tauber, Stefan IV-285
 Teixeira-Faria, Pedro M. I-278, I-309
 Tengku Wook, Tengku Siti Meriam IV-627
 Terai, Hitoshi IV-384
 Tews, Tessa-Karina III-577
 Tharanathan, Anand I-378
 Thiel, Simon III-474
 Thiruvengada, Hari I-378
 Tognetti, Simone IV-649
 Tokumaru, Masataka IV-121
 Tomimatsu, Kiyoshi I-45, III-30
 Trapp, Marcus I-238
 Trappeniers, Lieven I-229
 Traue, Harald C. III-603
 Troccaz, Jocelyne IV-202
 Tsai, Chih-Chieh I-82
 Tsai, Pei-Hsuan III-107
 Tsai, Pen-Yan I-82
 Tscherrig, Julien II-222
 Tseng, Wen Chieh II-491
 Tseng, Ya-Ying IV-570
 Tsianos, Nikos I-388
 Tsuchiya, Fumito I-588, I-628, IV-174, IV-210, IV-226

- Tsujino, Yoshihiro IV-351
 Tullis, Tom I-331
 Tulu, Bengisu IV-245
 Tung, Fang-Wu IV-637
 Turconi, Anna Carla I-406
 Uang, Shiaw-Tsyr IV-490
 Ullah, Sehat II-140
 Ur, Blase III-137
 Uster, Guillaume III-301
 Valls, Alícia I-136
 van Doesburg, Willem II-575
 Vasconcelos, Verónica III-293
 Vella, Frédéric IV-530
 Vera-Muñoz, Cecilia III-129, IV-219
 Vigouroux, Nadine IV-530
 Vijayaraghavan, P. IV-274
 Vladušić, Daniel III-519
 Vogel-Heuser, Birgit IV-297
 Wajima, Masayuki IV-617
 Walter, Steffen III-603
 Wan, Tao I-562
 Wan Adnan, Wan Adilah I-17
 Wang, Ai-Ling IV-107
 Wang, Cheng-Han III-311
 Wang, Hongmei III-338
 Wang, Liang I-660
 Wang, Mo I-142
 Wang, Ying III-346
 Wang, Yuxia IV-694
 Waselewsky, Marc IV-678
 Watanabe, Tomio II-585, II-603
 Watanabe, Yoko IV-413
 Wei, Ke III-220
 Werner, Günter I-618
 Wiley, Cyndi IV-285
 Wohlfarter, Martin II-125
 Wu, Andy II-148
 Wu, Chenjun I-507
 Wu, Hao I-580
 Wu, Zhongke I-555, I-670
 Wysk, Richard I-368
 Xie, Bin I-580
 Xie, Wenkui I-670
 Xu, Qingqing I-677
 Xu, Wenchang III-117
 Yajima, Hiroshi IV-113
 Yamaguchi, Takehiko II-3, II-62, II-158
 Yamaguchi, Takumi II-436
 Yamaguchi, Tatsuya IV-121
 Yamaguchi, Yoshihisa III-357
 Yamamoto, Michiya II-585, II-603
 Yamamoto, Tomohito I-517
 Yamanaka, Tsutomu IV-318
 Yamanishi, Yuya II-436
 Yang, Hsiao-Fang IV-77
 Yang, Mau-Tsuen IV-131
 Yasuda, Atsushi II-351
 Yeh, Ching-Yu IV-570
 Yen, Yi-Di III-185
 Yin, Jing I-414
 Yin, Qian I-598
 Yin, Xiupu I-684
 Yokokawa, Sho IV-237
 Yokoyama, Saya II-436
 Yoshimi, Hironori IV-226
 Yoshino, Takashi II-368, II-378
 Yu, Chun I-507
 Yu, Guo-Jhen III-250
 Yu, Lu III-346
 Yun, Myung Hwan IV-257
 Zablotskaya, Kseniya I-527
 Zablotskiy, Sergey I-527
 Zaki, Michael I-319
 Zaragozá, Irene I-423
 Zets, Gary III-321
 Zhan, Jia-Xuan III-259
 Zhang, Emily III-137
 Zhang, Huiting III-346
 Zhang, Kan III-346
 Zhang, Shaopeng I-142
 Zhang, Xian II-396
 Zhang, Xuhui I-450
 Zhao, Huiqin I-580
 Zhao, Qingjie IV-694
 Zheng, Xianjun Sam I-142
 Zheng, Xin I-677
 Zhou, Jinglin I-684
 Zhou, Mingquan I-555, I-670
 Zhou, Yun II-500
 Zhu, Haijiaing I-684
 Zhu, Miaoqi IV-659
 Zimmerman, Guy W. I-432