

Masaaki Kurosu (Ed.)

LNC8 8007

Human-Computer Interaction

Interaction Modalities and Techniques

15th International Conference, HCI International 2013
Las Vegas, NV, USA, July 2013
Proceedings, Part IV

4
Part IV



 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

Masaaki Kurosu (Ed.)

Human-Computer Interaction

Interaction Modalities and Techniques

15th International Conference, HCI International 2013
Las Vegas, NV, USA, July 21-26, 2013
Proceedings, Part IV



Springer

Volume Editor

Masaaki Kurosu

The Open University of Japan

2-11 Wakaba, Mihama-ku, Chiba-shi 261-8586, Japan

E-mail: masaakikurosu@spa.nifty.com

ISSN 0302-9743

e-ISSN 1611-3349

ISBN 978-3-642-39329-7

e-ISBN 978-3-642-39330-3

DOI 10.1007/978-3-642-39330-3

Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2013941394

CR Subject Classification (1998): H.5, I.2.7, I.2.9-11, K.4, H.3

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

© Springer-Verlag Berlin Heidelberg 2013

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

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

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

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 15th International Conference on Human–Computer Interaction, HCI International 2013, was held in Las Vegas, Nevada, USA, 21–26 July 2013, incorporating 12 conferences / thematic areas:

Thematic areas:

- Human–Computer Interaction
- Human Interface and the Management of Information

Affiliated conferences:

- 10th International Conference on Engineering Psychology and Cognitive Ergonomics
- 7th International Conference on Universal Access in Human–Computer Interaction
- 5th International Conference on Virtual, Augmented and Mixed Reality
- 5th International Conference on Cross-Cultural Design
- 5th International Conference on Online Communities and Social Computing
- 7th International Conference on Augmented Cognition
- 4th International Conference on Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management
- 2nd International Conference on Design, User Experience and Usability
- 1st International Conference on Distributed, Ambient and Pervasive Interactions
- 1st International Conference on Human Aspects of Information Security, Privacy and Trust

A total of 5210 individuals from academia, research institutes, industry and governmental agencies from 70 countries submitted contributions, and 1666 papers and 303 posters 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 Masaaki Kurosu, contains papers focusing on the thematic area of Human–Computer Interaction, and addressing the following major topics:

- Speech, Natural Language and Auditory Interfaces
- Gesture and Eye-Gaze-Based Interaction
- Touch-Based Interaction
- Haptic Interaction
- Graphical User Interfaces and Visualisation

The remaining volumes of the HCI International 2013 proceedings are:

- Volume 1, LNCS 8004, Human–Computer Interaction: Human-Centred Design Approaches, Methods, Tools and Environments (Part I), edited by Masaaki Kurosu
- Volume 2, LNCS 8005, Human–Computer Interaction: Applications and Services (Part II), edited by Masaaki Kurosu
- Volume 3, LNCS 8006, Human–Computer Interaction: Users and Contexts of Use (Part III), edited by Masaaki Kurosu
- Volume 5, LNCS 8008, Human–Computer Interaction: Towards Intelligent and Implicit Interaction (Part V), edited by Masaaki Kurosu
- Volume 6, LNCS 8009, Universal Access in Human–Computer Interaction: Design Methods, Tools and Interaction Techniques for eInclusion (Part I), edited by Constantine Stephanidis and Margherita Antona
- Volume 7, LNCS 8010, Universal Access in Human–Computer Interaction: User and Context Diversity (Part II), edited by Constantine Stephanidis and Margherita Antona
- Volume 8, LNCS 8011, Universal Access in Human–Computer Interaction: Applications and Services for Quality of Life (Part III), edited by Constantine Stephanidis and Margherita Antona
- Volume 9, LNCS 8012, Design, User Experience, and Usability: Design Philosophy, Methods and Tools (Part I), edited by Aaron Marcus
- Volume 10, LNCS 8013, Design, User Experience, and Usability: Health, Learning, Playing, Cultural, and Cross-Cultural User Experience (Part II), edited by Aaron Marcus
- Volume 11, LNCS 8014, Design, User Experience, and Usability: User Experience in Novel Technological Environments (Part III), edited by Aaron Marcus
- Volume 12, LNCS 8015, Design, User Experience, and Usability: Web, Mobile and Product Design (Part IV), edited by Aaron Marcus
- Volume 13, LNCS 8016, Human Interface and the Management of Information: Information and Interaction Design (Part I), edited by Sakae Yamamoto
- Volume 14, LNCS 8017, Human Interface and the Management of Information: Information and Interaction for Health, Safety, Mobility and Complex Environments (Part II), edited by Sakae Yamamoto
- Volume 15, LNCS 8018, Human Interface and the Management of Information: Information and Interaction for Learning, Culture, Collaboration and Business (Part III), edited by Sakae Yamamoto
- Volume 16, LNAI 8019, Engineering Psychology and Cognitive Ergonomics: Understanding Human Cognition (Part I), edited by Don Harris
- Volume 17, LNAI 8020, Engineering Psychology and Cognitive Ergonomics: Applications and Services (Part II), edited by Don Harris
- Volume 18, LNCS 8021, Virtual, Augmented and Mixed Reality: Designing and Developing Augmented and Virtual Environments (Part I), edited by Randall Shumaker
- Volume 19, LNCS 8022, Virtual, Augmented and Mixed Reality: Systems and Applications (Part II), edited by Randall Shumaker

- Volume 20, LNCS 8023, Cross-Cultural Design: Methods, Practice and Case Studies (Part I), edited by P.L. Patrick Rau
- Volume 21, LNCS 8024, Cross-Cultural Design: Cultural Differences in Everyday Life (Part II), edited by P.L. Patrick Rau
- Volume 22, LNCS 8025, Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management: Healthcare and Safety of the Environment and Transport (Part I), edited by Vincent G. Duffy
- Volume 23, LNCS 8026, Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management: Human Body Modeling and Ergonomics (Part II), edited by Vincent G. Duffy
- Volume 24, LNAI 8027, Foundations of Augmented Cognition, edited by Dylan D. Schmorrow and Cali M. Fidopiastis
- Volume 25, LNCS 8028, Distributed, Ambient and Pervasive Interactions, edited by Norbert Streitz and Constantine Stephanidis
- Volume 26, LNCS 8029, Online Communities and Social Computing, edited by A. Ant Ozok and Panayiotis Zaphiris
- Volume 27, LNCS 8030, Human Aspects of Information Security, Privacy and Trust, edited by Louis Marinos and Ioannis Askoxylakis
- Volume 28, CCIS 373, HCI International 2013 Posters Proceedings (Part I), edited by Constantine Stephanidis
- Volume 29, CCIS 374, HCI International 2013 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 affiliated conferences and thematic areas, listed below, for their contribution to the highest scientific quality and the overall success of the HCI International 2013 conference.

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

I would also like to thank for their contribution towards the smooth organization of the HCI International 2013 Conference the members of the Human-Computer Interaction Laboratory of ICS-FORTH, and in particular George Paparoulis, Maria Pitsoulaki, Stavroula Ntoa, Maria Bouhli and George Kapnas.

May 2013

Constantine Stephanidis
General Chair, HCI International 2013

Organization

Human–Computer Interaction

Program Chair: Masaaki Kurosu, Japan

Jose Abdelnour-Nocera, UK	Kyungdoh Kim, South Korea
Sebastiano Bagnara, Italy	Heidi Krömker, Germany
Simone Barbosa, Brazil	Chen Ling, USA
Tomas Berns, Sweden	Yan Liu, USA
Nigel Bevan, UK	Zhengjie Liu, P.R. China
Simone Borsci, UK	Loïc Martínez Normand, Spain
Apala Lahiri Chavan, India	Chang S. Nam, USA
Sherry Chen, Taiwan	Naoko Okuizumi, Japan
Kevin Clark, USA	Noriko Osaka, Japan
Torkil Clemmensen, Denmark	Philippe Palanque, France
Xiaowen Fang, USA	Hans Persson, Sweden
Shin'ichi Fukuzumi, Japan	Ling Rothrock, USA
Vicki Hanson, UK	Naoki Sakakibara, Japan
Ayako Hashizume, Japan	Dominique Scapin, France
Anzai Hiroyuki, Italy	Guangfeng Song, USA
Sheue-Ling Hwang, Taiwan	Sanjay Tripathi, India
Wonil Hwang, South Korea	Chui Yin Wong, Malaysia
Minna Isomursu, Finland	Toshiki Yamaoka, Japan
Yong Gu Ji, South Korea	Kazuhiko Yamazaki, Japan
Esther Jun, USA	Ryoji Yoshitake, Japan
Mitsuhiko Karashima, Japan	Silvia Zimmermann, Switzerland

Human Interface and the Management of Information

Program Chair: Sakae Yamamoto, Japan

Hans-Jorg Bullinger, Germany	Mark Lehto, USA
Alan Chan, Hong Kong	Hiroyuki Miki, Japan
Gilsoo Cho, South Korea	Hirohiko Mori, Japan
Jon R. Gunderson, USA	Fiona Fui-Hoon Nah, USA
Shin'ichi Fukuzumi, Japan	Shogo Nishida, Japan
Michitaka Hirose, Japan	Robert Proctor, USA
Jhilmil Jain, USA	Youngho Rhee, South Korea
Yasufumi Kume, Japan	Katsunori Shimohara, Japan

Michale Smith, USA
Tsutomu Tabe, Japan
Hiroshi Tsuji, Japan

Kim-Phuong Vu, USA
Tomio Watanabe, Japan
Hidekazu Yoshikawa, Japan

Engineering Psychology and Cognitive Ergonomics

Program Chair: Don Harris, UK

Guy Andre Boy, USA
Joakim Dahlman, Sweden
Trevor Dobbins, UK
Mike Feary, USA
Shan Fu, P.R. China
Michaela Heese, Austria
Hung-Sying Jing, Taiwan
Wen-Chin Li, Taiwan
Mark A. Neerinx, The Netherlands
Jan M. Noyes, UK
Taezoon Park, Singapore

Paul Salmon, Australia
Axel Schulte, Germany
Siraj Shaikh, UK
Sarah C. Sharples, UK
Anthony Smoker, UK
Neville A. Stanton, UK
Alex Stedmon, UK
Xianghong Sun, P.R. China
Andrew Thatcher, South Africa
Matthew J.W. Thomas, Australia
Rolf Zon, The Netherlands

Universal Access in Human–Computer Interaction

Program Chairs: Constantine Stephanidis, Greece, and Margherita Antona, Greece

Julio Abascal, Spain
Ray Adams, UK
Gisela Susanne Bahr, USA
Margit Betke, USA
Christian Bühler, Germany
Stefan Carmien, Spain
Jerzy Charytonowicz, Poland
Carlos Duarte, Portugal
Pier Luigi Emiliani, Italy
Qin Gao, P.R. China
Andrina Granić, Croatia
Andreas Holzinger, Austria
Josette Jones, USA
Simeon Keates, UK

Georgios Kouroupetroglou, Greece
Patrick Langdon, UK
Seongil Lee, Korea
Ana Isabel B.B. Paraguay, Brazil
Helen Petrie, UK
Michael Pieper, Germany
Enrico Pontelli, USA
Jaime Sanchez, Chile
Anthony Savidis, Greece
Christian Stary, Austria
Hirotada Ueda, Japan
Gerhard Weber, Germany
Harald Weber, Germany

Virtual, Augmented and Mixed Reality

Program Chair: Randall Shumaker, USA

Waymon Armstrong, USA
 Juan Cendan, USA
 Rudy Darken, USA
 Cali M. Fidopiastis, USA
 Charles Hughes, USA
 David Kaber, USA
 Hirokazu Kato, Japan
 Denis Laurendeau, Canada
 Fotis Liarokapis, UK

Mark Livingston, USA
 Michael Macedonia, USA
 Gordon Mair, UK
 Jose San Martin, Spain
 Jacquelyn Morie, USA
 Albert “Skip” Rizzo, USA
 Kay Stanney, USA
 Christopher Stapleton, USA
 Gregory Welch, USA

Cross-Cultural Design

Program Chair: P.L. Patrick Rau, P.R. China

Pilsung Choe, P.R. China
 Henry Been-Lirn Duh, Singapore
 Vanessa Evers, The Netherlands
 Paul Fu, USA
 Zhiyong Fu, P.R. China
 Fu Guo, P.R. China
 Sung H. Han, Korea
 Toshikazu Kato, Japan
 Dyi-Yih Michael Lin, Taiwan
 Rungtai Lin, Taiwan

Sheau-Farn Max Liang, Taiwan
 Liang Ma, P.R. China
 Alexander Mädche, Germany
 Katsuhiko Ogawa, Japan
 Tom Plocher, USA
 Kerstin Röse, Germany
 Supriya Singh, Australia
 Hsiu-Ping Yueh, Taiwan
 Liang (Leon) Zeng, USA
 Chen Zhao, USA

Online Communities and Social Computing

Program Chairs: A. Ant Ozok, USA, and Panayiotis Zaphiris, Cyprus

Areej Al-Wabil, Saudi Arabia
 Leonelo Almeida, Brazil
 Bjørn Andersen, Norway
 Chee Siang Ang, UK
 Aneesha Bakharia, Australia
 Ania Bobrowicz, UK
 Paul Cairns, UK
 Farzin Deravi, UK
 Andri Ioannou, Cyprus
 Slava Kisilevich, Germany

Niki Lambropoulos, Greece
 Effie Law, Switzerland
 Soo Ling Lim, UK
 Fernando Loizides, Cyprus
 Gabriele Meiselwitz, USA
 Anthony Norcio, USA
 Elaine Raybourn, USA
 Panote Siriaraya, UK
 David Stuart, UK
 June Wei, USA

Augmented Cognition

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

Robert Arrabito, Canada

Richard Backs, USA

Chris Berka, USA

Joseph Cohn, USA

Martha E. Crosby, USA

Julie Drexler, USA

Ivy Estabrooke, USA

Chris Forsythe, USA

Wai Tat Fu, USA

Rodolphe Gentili, USA

Marc Grootjen, The Netherlands

Jefferson Grubb, USA

Ming Hou, Canada

Santosh Mathan, USA

Rob Matthews, Australia

Dennis McBride, USA

Jeff Morrison, USA

Mark A. Neerincx, The Netherlands

Denise Nicholson, USA

Banu Onaral, USA

Lee Sciarini, USA

Kay Stanney, USA

Roy Stripling, USA

Rob Taylor, UK

Karl van Orden, USA

Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management

Program Chair: Vincent G. Duffy, USA and Russia

Karim Abdel-Malek, USA

Giuseppe Andreoni, Italy

Daniel Carruth, USA

Eliza Yingzi Du, USA

Enda Fallon, Ireland

Afzal Godil, USA

Ravindra Goonetilleke, Hong Kong

Bo Hoege, Germany

Waldemar Karwowski, USA

Zhizhong Li, P.R. China

Kang Li, USA

Tim Marler, USA

Michelle Robertson, USA

Matthias Rötting, Germany

Peter Vink, The Netherlands

Mao-Jiun Wang, Taiwan

Xuguang Wang, France

Jingzhou (James) Yang, USA

Xiangan Yuan, P.R. China

Gülcin Yücel Hoge, Germany

Design, User Experience, and Usability

Program Chair: Aaron Marcus, USA

Sisira Adikari, Australia

Ronald Baecker, Canada

Arne Berger, Germany

Jamie Blustein, Canada

Ana Boa-Ventura, USA

Jan Brejcha, Czech Republic

Lorenzo Cantoni, Switzerland

Maximilian Eibl, Germany

Anthony Faiola, USA
 Emilie Gould, USA
 Zelda Harrison, USA
 Rüdiger Heimgärtner, Germany
 Brigitte Herrmann, Germany
 Steffen Hess, Germany
 Kaleem Khan, Canada

Jennifer McGinn, USA
 Francisco Rebelo, Portugal
 Michael Renner, Switzerland
 Kerem Rızvanoğlu, Turkey
 Marcelo Soares, Brazil
 Christian Sturm, Germany
 Michele Visciola, Italy

Distributed, Ambient and Pervasive Interactions

Program Chairs: Norbert Streitz, Germany, and Constantine Stephanidis, Greece

Emile Aarts, The Netherlands
 Adnan Abu-Dayya, Qatar
 Juan Carlos Augusto, UK
 Boris de Ruyter, The Netherlands
 Anind Dey, USA
 Dimitris Grammenos, Greece
 Nuno M. Guimaraes, Portugal
 Shin'ichi Konomi, Japan
 Carsten Magerkurth, Switzerland

Christian Müller-Tomfelde, Australia
 Fabio Paternó, Italy
 Gilles Privat, France
 Harald Reiterer, Germany
 Carsten Röcker, Germany
 Reiner Wichert, Germany
 Woontack Woo, South Korea
 Xenophon Zabulis, Greece

Human Aspects of Information Security, Privacy and Trust

Program Chairs: Louis Marinos, ENISA EU, and Ioannis Askoxylakis, Greece

Claudio Agostino Ardagna, Italy
 Zinaida Benenson, Germany
 Daniele Catteddu, Italy
 Raoul Chiesa, Italy
 Bryan Cline, USA
 Sadie Creese, UK
 Jorge Cuellar, Germany
 Marc Dacier, USA
 Dieter Gollmann, Germany
 Kirstie Hawkey, Canada
 Jaap-Henk Hoepman, The Netherlands
 Cagatay Karabat, Turkey
 Angelos Keromytis, USA
 Ayako Komatsu, Japan

Ronald Leenes, The Netherlands
 Javier Lopez, Spain
 Steve Marsh, Canada
 Gregorio Martinez, Spain
 Emilio Mordini, Italy
 Yuko Murayama, Japan
 Masakatsu Nishigaki, Japan
 Aljosa Pasic, Spain
 Milan Petković, The Netherlands
 Joachim Posegga, Germany
 Jean-Jacques Quisquater, Belgium
 Damien Sauveron, France
 George Spanoudakis, UK
 Kerry-Lynn Thomson, South Africa

Julien Touzeau, France
Theo Tryfonas, UK
João Vilela, Portugal

Claire Vishik, UK
Melanie Volkamer, Germany

External Reviewers

Maysoon Abulhair, Saudi Arabia
Ilia Adami, Greece
Vishal Barot, UK
Stephan Böhm, Germany
Vassilis Charissis, UK
Francisco Cipolla-Ficarra, Spain
Maria De Marsico, Italy
Marc Fabri, UK
David Fonseca, Spain
Linda Harley, USA
Yasushi Ikei, Japan
Wei Ji, USA
Nouf Khashman, Canada
John Killilea, USA
Iosif Klironomos, Greece
Ute Klotz, Switzerland
Maria Korozi, Greece
Kentaro Kotani, Japan

Vassilis Kouroumalis, Greece
Stephanie Lackey, USA
Janelle LaMarche, USA
Asterios Leonidis, Greece
Nickolas Macchiarella, USA
George Margetis, Greece
Matthew Marraffino, USA
Joseph Mercado, USA
Claudia Mont'Alvão, Brazil
Yoichi Motomura, Japan
Karsten Nebe, Germany
Stavroula Ntoa, Greece
Martin Osen, Austria
Stephen Prior, UK
Farid Shirazi, Canada
Jan Stelovsky, USA
Sarah Swierenga, USA

HCI International 2014

The 16th International Conference on Human–Computer Interaction, HCI International 2014, will be held jointly with the affiliated conferences in the summer of 2014. It will cover a broad spectrum of themes related to Human–Computer Interaction, including theoretical issues, methods, tools, processes and case studies in HCI design, as well as novel interaction techniques, interfaces and applications. The proceedings will be published by Springer. More information 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 IV

Speech, Natural Language and Auditory Interfaces

Controlling Interaction in Multilingual Conversation	3
<i>Christina Alexandris</i>	
Linguistic Processing of Implied Information and Connotative Features in Multilingual HCI Applications	13
<i>Christina Alexandris and Ioanna Malagardi</i>	
Investigating the Impact of Combining Speech and Earcons to Communicate Information in E-government Interfaces	23
<i>Dimitrios Rigas and Badr Almutairi</i>	
Evaluation of WikiTalk – User Studies of Human-Robot Interaction	32
<i>Dimitra Anastasiou, Kristiina Jokinen, and Graham Wilcock</i>	
Robust Multi-Modal Speech Recognition in Two Languages Utilizing Video and Distance Information from the Kinect	43
<i>Georgios Galatas, Gerasimos Potamianos, and Fillia Makedon</i>	
The Ecological AUI (Auditory User Interface) Design and Evaluation of User Acceptance for Various Tasks on Smartphones	49
<i>Myounghoon Jeon and Ju-Hwan Lee</i>	
Speech-Based Text Correction Patterns in Noisy Environment	59
<i>Ladislav Kunc, Tomáš Macek, Martin Labský, and Jan Kleindienst</i>	
Multimodal Smart Interactive Presentation System	67
<i>Hoang-An Le, Khoi-Nguyen C. Mac, Truong-An Pham, Vinh-Tiep Nguyen, and Minh-Triet Tran</i>	
Multimodal Mathematical Expressions Recognition: Case of Speech and Handwriting	77
<i>Sofiane Medjkoune, Harold Mouchere, Simon Petitrenaud, and Christian Viard-Gaudin</i>	
‘Realness’ in Chatbots: Establishing Quantifiable Criteria	87
<i>Kellie Morrissey and Jurek Kirakowski</i>	
Grounding and Turn-Taking in Multimodal Multiparty Conversation . . .	97
<i>David Novick and Iván Gris</i>	
Situated Multiparty Interaction between Humans and Agents	107
<i>Aasish Pappu, Ming Sun, Seshadri Sridharan, and Alex Rudnicky</i>	

Enhancing Human Computer Interaction with Episodic Memory in a Virtual Guide	117
<i>Felix Rabe and Ipke Wachsmuth</i>	
System of Generating Japanese Sound Symbolic Expressions Using Genetic Algorithm	126
<i>Yuichiro Shimizu, Tetsuaki Nakamura, and Maki Sakamoto</i>	
A Knowledge Elicitation Study for Collaborative Dialogue Strategies Used to Handle Uncertainties in Speech Communication While Using GIS	135
<i>Hongmei Wang, Ava Gailliot, Douglas Hyden, and Ryan Lietzenmayer</i>	

Gesture and Eye-Gaze Based Interaction

Context-Based Bounding Volume Morphing in Pointing Gesture Application	147
<i>Andreas Braun, Arthur Fischer, Alexander Marinc, Carsten Stockl�w, and Martin Majewski</i>	
Gesture vs. Gesticulation: A Test Protocol	157
<i>Francesco Carrino, Antonio Ridi, Rolf Ingold, Omar Abou Khaled, and Elena Mugellini</i>	
Functional Gestures for Human-Environment Interaction	167
<i>Stefano Carrino, Maurizio Caon, Omar Abou Khaled, Rolf Ingold, and Elena Mugellini</i>	
A Dynamic Fitting Room Based on Microsoft Kinect and Augmented Reality Technologies	177
<i>Hsien-Tsung Chang, Yu-Wen Li, Huan-Ting Chen, Shih-Yi Feng, and Tsung-Tien Chien</i>	
Gesture-Based Applications for Elderly People	186
<i>Weiqin Chen</i>	
MOBAJES: Multi-user Gesture Interaction System with Wearable Mobile Device	196
<i>Enkhbat Davaasuren and Jiro Tanaka</i>	
Head-Free, Remote Gaze Detection System Based on Pupil-Corneal Reflection Method with Using Two Video Cameras – One-Point and Nonlinear Calibrations	205
<i>Yoshinobu Ebisawa and Kiyotaka Fukumoto</i>	
Design and Usability Analysis of Gesture-Based Control for Common Desktop Tasks	215
<i>Farzin Farhadi-Niaki, S. Ali Etemad, and Ali Arya</i>	

Study of Eye-Glance Input Interface	225
<i>Dekun Gao, Naoaki Itakura, Tota Mizuno, and Kazuyuki Mito</i>	
Multi-User Interaction with Shadows	235
<i>Tomomi Gotoh, Takahiro Kida, Munehiro Takimoto, and Yasushi Kambayashi</i>	
Intent Capturing through Multimodal Inputs	243
<i>Weimin Guo, Cheng Cheng, Mingkai Cheng, Yonghan Jiang, and Honglin Tang</i>	
Robust Hand Tracking in Realtime Using a Single Head-Mounted RGB Camera	252
<i>Jan Hendrik Hammer and Jürgen Beyerer</i>	
Multimodal Feedback in First Encounter Interactions	262
<i>Kristiina Jokinen</i>	
Keyboard Clawing: Input Method by Clawing Key Tops	272
<i>Toshifumi Kurosawa, Buntarou Shizuki, and Jiro Tanaka</i>	
Finger Controller: Natural User Interaction Using Finger Gestures	281
<i>Unseok Lee and Jiro Tanaka</i>	
A Method for Single Hand Fist Gesture Input to Enhance Human Computer Interaction	291
<i>Tao Ma, William Wee, Chia Yung Han, and Xuefu Zhou</i>	
Kinect© as Interaction Device with a Tiled Display	301
<i>Amilcar Meneses Viveros and Erika Hernández Rubio</i>	
Study on Cursor Shape Suitable for Eye-gaze Input System	312
<i>Atsuo Murata, Raku Uetsugi, and Takehito Hayami</i>	
Study on Character Input Methods Using Eye-gaze Input Interface	320
<i>Atsuo Murata, Kazuya Hayashi, Makoto Moriwaka, and Takehito Hayami</i>	
Proposal of Estimation Method of Stable Fixation Points for Eye-gaze Input Interface	330
<i>Atsuo Murata, Takehito Hayami, and Keita Ochi</i>	
Modeling Situation-Dependent Nonverbal Expressions for a Pair of Embodied Agent in a Dialogue Based on Conversations in TV Programs	340
<i>Keita Okuwuchi, Koh Kakusho, Takatsugu Kojima, and Daisuke Katagami</i>	
Research on a Large Digital Desktop Integrated in a Traditional Environment for Informal Collaboration	348
<i>Mariano Perez Pelaez, Ryo Suzuki, and Ikuro Choh</i>	

Using Kinect for 2D and 3D Pointing Tasks: Performance Evaluation ...	358
<i>Alexandros Pino, Evangelos Tzemis, Nikolaos Ioannou, and Georgios Kouroupetroglou</i>	
Conditions of Applications, Situations and Functions Applicable to Gesture Interface	368
<i>Taebeum Ryu, Jaehong Lee, Myung Hwan Yun, and Ji Hyoun Lim</i>	
Communication Analysis of Remote Collaboration System with Arm Scaling Function	378
<i>Nobuchika Sakata, Tomoyuki Kobayashi, and Shogo Nishida</i>	
Two Handed Mid-Air Gestural HCI: Point + Command	388
<i>Matthias Schwaller, Simon Brunner, and Denis Lalanne</i>	
Experimental Study Toward Modeling of the Uncanny Valley Based on Eye Movements on Human/Non-human Faces	398
<i>Yoshimasa Tawatsuji, Kazuaki Kojima, and Tatsunori Matsui</i>	
Multi-party Human-Machine Interaction Using a Smart Multimodal Digital Signage	408
<i>Tony Tung, Randy Gomez, Tatsuya Kawahara, and Takashi Matsuyama</i>	
A Remote Pointing Technique Using Pull-out	416
<i>Takuto Yoshikawa, Yuusaku Mita, Takuro Kuribara, Buntarou Shizuki, and Jiro Tanaka</i>	

Touch-Based Interaction

Human Centered Design Approach to Integrate Touch Screen in Future Aircraft Cockpits	429
<i>Jérôme Barbé, Marion Wolff, and Régis Mollard</i>	
Evaluating Devices and Navigation Tools in 3D Environments	439
<i>Marcela Câmara, Priscilla Fonseca de Abreu Braz, Ingrid Monteiro, Alberto Raposo, and Simone Diniz Junqueira Barbosa</i>	
Computational Cognitive Modeling of Touch and Gesture on Mobile Multitouch Devices: Applications and Challenges for Existing Theory ...	449
<i>Kristen K. Greene, Franklin P. Tamborello, and Ross J. Micheals</i>	
A Page Navigation Technique for Overlooking Content in a Digital Magazine	456
<i>Yuichiro Kinoshita, Masayuki Sugiyama, and Kentaro Go</i>	

Effect of Unresponsive Time for User's Touch Action of Selecting an Icon on the Video Mirror Interface	462
<i>Kazuyoshi Murata, Masatsugu Hattori, and Yu Shibuya</i>	
Evaluation of a Soft-Surfaced Multi-touch Interface	469
<i>Anna Noguchi, Toshifumi Kurosawa, Ayaka Suzuki, Yuichiro Sakamoto, Tatsuhito Oe, Takuto Yoshikawa, Buntarou Shizuki, and Jiro Tanaka</i>	
Recognition of Multi-touch Drawn Sketches	479
<i>Michael Schmidt and Gerhard Weber</i>	
A Web Browsing Method on Handheld Touch Screen Devices for Preventing from Tapping Unintended Links	491
<i>Yu Shibuya, Hikaru Kawakatsu, and Kazuyoshi Murata</i>	
Real Time Mono-vision Based Customizable Virtual Keyboard Using Finger Tip Speed Analysis	497
<i>Sumit Srivastava and Ramesh Chandra Tripathi</i>	
Human Factor Research of User Interface for 3D Display	506
<i>Chih-Hung Ting, Teng-Yao Tsai, Yi-Pai Huang, Wen-Jun Zeng, and Ming-Hui Lin</i>	
Collaborative Smart Virtual Keyboard with Word Predicting Function	513
<i>Chau Thai Truong, Duy-Hung Nguyen-Huynh, Minh-Triet Tran, and Anh-Duc Duong</i>	
The Implementation of Multi-touch Table to Support the Military Decision Making through Critical Success Factors (CSFs)	523
<i>Norshahriah Wahab and Halimah Badioze Zaman</i>	
Design of a Visual Query Language for Geographic Information System on a Touch Screen	530
<i>Siju Wu, Samir Otmame, Guillaume Moreau, and Myriam Servières</i>	
Target Orientation Effects on Movement Time in Rapid Aiming Tasks	540
<i>Yugang Zhang, Bifeng Song, and Wensheng Min</i>	

Haptic Interaction

Comparison of Enhanced Visual and Haptic Features in a Virtual Reality-Based Haptic Simulation	551
<i>Michael Clamann, Wenqi Ma, and David B. Kaber</i>	

Influence of Haptic Feedback on a Pointing Task in a Haptically Enhanced 3D Virtual Environment	561
<i>Brendan Corbett, Takehiko Yamaguchi, Shijing Liu, Lixiao Huang, Sangwoo Bahn, and Chang S. Nam</i>	
Design of a Wearable Haptic Vest as a Supportive Tool for Navigation	568
<i>Anak Agung Gede Dharma, Takuma Oami, Yuhki Obata, Li Yan, and Kiyoshi Tomimatsu</i>	
Mapping Texture Phase Diagram of Artificial Haptic Stimuli Generated by Vibrotactile Actuators	578
<i>Anak Agung Gede Dharma and Kiyoshi Tomimatsu</i>	
Preliminary Design of Haptic Icons from Users	587
<i>Wonil Hwang and Dongsoo Kim</i>	
Assessing the Effectiveness of Vibrotactile Feedback on a 2D Navigation Task	594
<i>Wooram Jeon, Yueqing Li, Sangwoo Bahn, and Chang S. Nam</i>	
Magnetic Field Based Near Surface Haptic and Pointing Interface	601
<i>Kasun Karunanayaka, Sanath Siriwardana, Chamari Edirisinghe, Ryohei Nakatsu, and Ponnampalam Gopalakrishnakone</i>	
Use of Reference Frame in Haptic Virtual Environments: Implications for Users with Visual Impairments	610
<i>Ja Young Lee, Sangwoo Bahn, and Chang S. Nam</i>	
Behavioral Characteristics of Users with Visual Impairment in Haptically Enhanced Virtual Environments	618
<i>Shijing Liu, Sangwoo Bahn, Heesun Choi, and Chang S. Nam</i>	
Graphical User Interfaces and Visualisation	
A Situation Awareness Assistant for Human Deep Space Exploration . . .	629
<i>Guy Andre Boy and Donald Platt</i>	
My-World-in-My-Tablet: An Architecture for People with Physical Impairment	637
<i>Mario Caruso, Febo Cincotti, Francesco Leotta, Massimo Mecella, Angela Riccio, Francesca Schettini, Luca Simione, and Tiziana Catarci</i>	
AHPM as a Proposal to Improve Interaction with Air Traffic Controllers	648
<i>Leonardo L.B.V. Cruciol and Li Weigang</i>	

Decision Space Visualization: Lessons Learned and Design Principles ...	658
<i>Jill L. Drury, Mark S. Pfaff, Gary L. Klein, and Yikun Liu</i>	
The Language of Motion: A Taxonomy for Interface	668
<i>Elaine Froehlich, Brian Lucid, and Heather Shaw</i>	
Adaptive Consoles for Supervisory Control of Multiple Unmanned Aerial Vehicles	678
<i>Christian Fuchs, Sérgio Ferreira, João Sousa, and Gil Gonçalves</i>	
A Web-Based Interface for a System That Designs Sensor Networks	688
<i>Lawrence J. Henschen and Julia C. Lee</i>	
An Interaction Concept for Public Displays and Mobile Devices in Public Transport	698
<i>Romina Kühn, Diana Lemme, and Thomas Schlegel</i>	
Study of Interaction Concepts in 3D Virtual Environment	706
<i>Vera Oblaender and Maximilian Eibl</i>	
Undo/Redo by Trajectory	712
<i>Tatsuhito Oe, Buntarou Shizuki, and Jiro Tanaka</i>	
Multi-layer Control and Graphical Feature Editing Using Server-Side Rendering on Ajax-GIS	722
<i>Takeo Sakairi, Takashi Tamada, Katsuyuki Kamei, and Yukio Goto</i>	
A Method for Discussing Musical Expression between Music Ensemble Players Using a Web-Based System	730
<i>Takehiko Sakamoto, Shin Takahashi, and Jiro Tanaka</i>	
A Study on Document Retrieval System Based on Visualization to Manage OCR Documents	740
<i>Kazuki Tamura, Tomohiro Yoshikawa, and Takeshi Furuhashi</i>	
Audio-Visual Documentation Method for Digital Storytelling for a Multimedia Art Project	750
<i>Chui Yin Wong, Chee Weng Khong, Kimberly Chu, Muhammad Asyraf Mhd Pauzi, and Man Leong Wong</i>	
Author Index	759

Part I
Speech, Natural Language and
Auditory Interfaces

Controlling Interaction in Multilingual Conversation

Christina Alexandris

National University of Athens, Greece
calexandris@gs.uoa.gr

Abstract. The present approach targets to provide a framework for facilitating multilingual interaction in online business meetings with an agenda as well as in similar applications in the service sector where there is a less task-oriented form of interaction. A basic problem to be addressed is the control of the topics covered during the interaction and the expression of opinion. In the proposed template-based approach, the System is proposed to act as a mediator to control the dialog flow, within the modeled framework of the sublanguage-specific and pragmatically related design.

Keywords: Templates, Simple Interlinguas, Non Task-related Speech Acts, Skype, subtitles.

1 Introduction

The domain of the proposed framework concerns routine business meetings via Skype with an agenda, namely standard, controlled conversation. This domain includes services requiring a less task-oriented form of interaction as well as statement of sentiment or opinion. The proposed framework is not recommended for business meetings concerning negotiations or business deals. Conditions involve multilingual conversation with speakers of less spoken languages and an average or less than average knowledge of English. The interaction involves Skype meetings with or without translated subtitles. Subtitles (text messages) are combined with spoken input. Users may speak at the same time while the corresponding, similar or even different text message is generated. The communicating parties have access to prosodic and paralinguistic information, such as tone of voice, as well as facial expressions and other paralinguistic elements.

The flow of the conversation can be checked by the System or the User, as well as the topics covered. This is achieved by intervening messages by the System, appearing in the screen of the interface. Furthermore, additional free input from the User's utterances may be processed after the interaction. Spoken interaction is processed with the use of Interlinguas (we propose the use of "Simple Interlinguas" – SILTs) [8] generated simultaneously with translated text messages chosen by the User.

The proposed service may be regarded as an alternative to email in routine business meetings, allowing face-to-face interaction and feed-back from paralinguistic elements, such as gestures, facial expression and tone of voice. Furthermore, in multilingual applications involving communication with a standard agenda, the proposed approach may be adaptable and reusable in respect to several languages.

Specifically, the present design concerns a Speech Act based template and agenda. This agenda may be visible or invisible to Users. The interaction takes place in a Directed Dialog [15][16] like form and related to the respective Speech Act. The template contains a set of sublanguage-specific questions, statements and answers, including opinions, some pre-existing, while others are left to be prepared by the User prior to the meeting. The template-agenda contains the topics covered during the interaction and reminds Users, if a topic is not covered. In other words, interaction is controlled by the templates of the System, which acts as a mediator.

2 Design Principles and Previous Approaches

Three factors may be taken into account for the present design. First, Service- Oriented Dialog Systems targeted towards the broad public involve a higher percentage of non-sublanguage specific vocabulary and a lower percentage of terminology and professional jargon. In particular, applications related to business meetings often involve expressions related to emotion and the statement of opinion (1). Second, unlike highly specialized Task-related Dialog Systems, in Service- Oriented Dialog Systems, the Human-Computer interaction taking place is directed towards two equally significant goals, namely the successful performance of the activated or requested task as well as User satisfaction and User-friendliness (2). These goals are related to requirements on the Satisfaction Level in respect to a System's evaluation criteria, namely perceived task success, comparability of human partner and trustworthiness [10]. It should be noted that the more goals to be achieved, the more parameters in the System Design and System Requirements, and, subsequently, Dialog Design are to be considered [14]. The diversity of a multilingual User Group (3) constitutes an additional factor in the present design, ranging from Users belonging to communities where any real experience with computers and electronic devices is restricted to only a minority of Users, to User Groups with an absolute familiarity of society with computers and electronic devices, including cases where a tendency towards attachment [9] or animism of these objects is observed.

2.1 Directed Dialogs

For achieving User-friendliness in multilingual Dialog Systems with a diversity of Users, strategies such as the use of Directed Dialogs [15] [16] using keywords offer a predefined pattern of User interaction with the System in order to prevent an uncontrolled number of possible forms and variations [8] in (a) the expression in each language and in (b) User behavior due to cultural and social factors.

The use of Directed Dialogs and Yes-No questions aims to the highest possible recognition rate of a very broad and varied user group and the use of free spoken input processes the detailed information involved in a complex application. Keyword recognition largely occurs within a Yes-No question sequence of a Directed Dialog.

Within the Directed-Dialog framework of Dialog Systems in the service sector, such as in the present application, additional types of Speech Acts are detected, other than strictly Task-related Speech Acts. These Non-Task-related Speech Acts, [2]

whose determination was based on data from European Union Projects [19][20][21][22] are used for tasks such as “Offer”, “Reminder” or “Manage-Waiting-Time” [2], mostly in messages generated by the System.

2.2 Template Generation

The present approach is based on previous practices concerning the use of templates interacting within a Directed Dialog framework. A typical Dialog System involving the use of templates is the CitizenShield Dialog System for consumer complaints [3][11] handling routine tasks involving food and manufactured products (namely complaints involving quality, product labels, prices etc.). The spoken input is automatically entered into templates containing a number of fields related to the categories and types of information concerning the product involved. Free spoken input is recorded within a defined period of time, following a question requiring detailed information and/or detailed descriptions. All spoken input, whether constituting an answer to a Yes-No question or constituting an answer to a question triggering a Free-Input answer, is automatically directed to the respective templates of a complaint form, which are filled in by the spoken utterances, recognized by the System’s Automatic Speech Recognition (ASR) component. The automatic filling-in of complaint forms with spoken input via the consumer organization’s call center is especially helpful to mobile Users and Users that have no internet access. The information contained in each field of the complaint form is automatically or manually processable, according to the type of task to be executed [3][11]. The generation of the template-based complaint forms is also aimed towards the construction of continually updated databases from which statistical and other types of information is retrievable for the use of companies, organizations or authorities or other interested parties [3][11].

In other words, in previous approaches, the templates are both registering and controlling User input [11]. In the present design, the System may also use the templates to check if all issues to be addressed are covered. In other words, the template also behaves as an agenda during the interaction. In addition, according to the indications of the template-agenda, the System may automatically intervene with asking additional questions, until the issue in question is addressed. Most of these questions are of the Non-task related Speech Act type, such as “Offer”, “Reminder” or “Manage-Waiting-Time” [2].

Furthermore, the present approach involves the activation of prepared answers contained in the template activated by the User in the appropriate step in the dialog.

2.3 Directed Dialogs and Interlinguas

For Multilingual Dialog Systems using Directed Dialogs, Simple Interlinguas (S-ILTs) are proposed (Table 1) [8], constituting lexical-based alternatives and a simplified form of Interlinguas. S-ILTs may be characterized by a very simple structure and with a weakened “frame” function [8] which in traditional Interlinguas summarizes the semantic content of a spoken utterance [7][13]. In S-ILTs, the semantic content of a spoken utterance is signaled by the respective topic of step in the dialog structure. Specifically, the

use of Directed Dialogs, and the Speech Acts performed in the dialog structure, the proposed lexical-based alternatives are linked to Speech Act types in respect to the steps in dialog context. Thus, the role of the “frame” is weakened and the role of dialog structure is reinforced. With the use of Directed Dialogs in multilingual applications, the proposed “Simple Interlinguas” (S-ILTs) [8] allow the recognition and isolated processing of keywords at the lexical level, a feature which facilitates the compatibility with Systems that support multiple languages and cases of polysemy and multiple grammatical functions of word types. Furthermore, Simple Interlinguas operating on the lexical level may also be directly linked to more language-independent entities such as the Universal Words (UWs) of the UNDL System of the United Nations [5][23] which are linked to each other by semantic relations such as hyperonymy and synonymy.

Table 1. Example of a Simple Interlingua (S-ILT) connected to the step in dialog structure and respective Speech Act for applications concerning the expression of opinion. Optional entries at Keyword Content level are functions “Who”, and “When” and respective lexical entries.

SPEECH ACT TOPIC OF STEP IN DIALOG STRUCTURE (SYSTEM):	S-ILT (RESPONSE FROM USER CONTAINING LANGUAGE-SPECIFIC EXPRESSIONS)
TOPIC-OF-STEP {REFUSE} =>	=> [S-FRAME TOPIC-OF-STEP {REFUSE} YES/NO (no-answer/no-expression) <i>Optional</i> : WHO(person), WHEN(time)]

However, it should be stressed that the proposed S-ILTs are sublanguage-specific and are modeled according to the Speech Acts in the steps of the dialog structure and keywords of the application domain. In other words, a rigid and controlled nature of a System’s dialog structure allows the successful function of the proposed S-Interlinguas, specifically, Directed Dialogs [15][16], involving Yes-No Questions or questions directed towards Keyword answers with related signaled topic (TOPIC-OF-STEP). The combination of the use of Directed Dialogs and S-Interlinguas allows the control and successful handling of a varied or ambiguous input, in accordance to the criteria of the Utterance Level (Question-Answer-Level) and the Functional Level, especially informativeness and intelligibility (Utterance Level) and the ease of use, interaction control and processing speed/smoothness (Functional Level) [10].

It should additionally be noted that the proposed S-ILTs allow minimum interference of language-specific factors, since they are designed to function within a very restricted sublanguage related to a specific task. They can, therefore, be adapted for a very diverse range of languages and language families, for example, Hindi and Chinese. This S-ILT framework is proposed for the present approach, however, any other Interlingua type, if appropriate for the languages concerned, may be used.

3 Interaction

In the present approach, interaction occurs in three levels in respect to interaction type (A) and also in three levels in respect to the chronological process of interaction (B).

The type of interaction (A) concerns the multimodal interaction with spoken utterances with Speech-to-Speech Translation with Interlinguas or S-ILTs (1) which occurs simultaneously with interaction with prepared texts appearing on the screen as subtitles during interaction with Skype (2) or any other form of visual interaction. These texts can be translated by a Text-to-Text translation System. The third type of interaction involves the production of filled-in templates (3) with the issues covered during the multimodal interaction process or pending issues, as well as additional information and comments produced by speakers in the form of free input. The proposed approach is strictly sublanguage and domain-specific, however, it is designed to be compatible with online (written text) Machine Translation systems, such as Google Translate, processing many languages. However, a language-specific choice is necessary for the spoken Machine Translation component in regard to Interlinguas (ILTs) or the proposed Simple Interlinguas (S-ILTs). The design of the proposed approach in respect to interaction type may be depicted in the following table (Table 2).

Table 2. Interaction Type: Input and Output

Input Type (During Interaction):	
SCREEN (e.g. Skype)	Template-Agenda ☒ [open/close] (3) <ul style="list-style-type: none"> •Registers Free Input •Checks topics covered •Generates messages
<i>Paralinguistic Data (Prosody, Gestures etc.)</i>	
(1) INTERLINGUAS (ILT OR S-ILT)	
Speech-to-Speech Translation	
(2) SUBTITLES	
Text-to-Text Translation (<i>e.g. Google Translate</i>)	
(3) SYSTEM MESSAGES TO USER	

Regarding the chronological process of interaction (B), the involvement of the Users of both ends concerns the time span prior to the actual interaction (I), the time span of the actual interaction (II) and the time span after the interaction (III).

3.1 Template-Agenda and Preparation of Interaction (I)

In the time span prior to the actual interaction (I), the Users prepare the set of questions, statements and possible set of answers within the framework of the template containing the types of information handled during interaction.

The prepared restricted set of questions, statements and set of possible answers are activated by the User during the time of the actual online interaction. The preparation of topics, agenda and general outline of dialog structure is determined by the type, content and style of routine communication defined according to the tasks and policy of the company or organization. In other words, the sublanguage-specific set of questions, statements and set of possible answers prepared by the Users is also determined by the tasks and policy of the company or organization. This template-agenda contains a set of sublanguage-specific questions, statements and answers, some pre-existing in the sublanguage-specific design, while others are left to be prepared by

the User prior to the meeting. However, the messages to be prepared by the User receive specific content tags related to the sublanguage of the template and the related Interlinguas and must contain words related to them and there are restrictions in respect to the length of the utterances used. In addition, the User may choose from a set of symbolic markers to indicate attitude or emotion, if desired or if necessary, for example “assertive” or “polite”. The prepared answers and other messages contained in the template are activated by the User in the appropriate step in the dialog. Each activated answer in the source language is automatically translated by a conventional Text-to-Text Machine Translation System and appears on the screen of the Receiver in the target language, while at the same time the User in the source language may utter the same or a similar sentence to the utterance activated. As an optional element, any symbolic markers indicating attitude or emotion may also be displayed on the screen of the Receiver. The template-agenda contains the topics covered during the interaction and reminds Users, if a topic is not covered. A similar process is activated if opinions are requested. The general outline and content of the template-agenda may be depicted in the following table (Table 3).

Table 3. Outline and content of the template-agenda

Prepared Utterances by User:	System:	Template-Agenda
ISSUE -1 [proposal]: Utterance 1 Utterance 2	ISSUE -1: CHECKED	∅
ISSUE -2 [check]: Utterance 1 Utterance 2	ISSUE -2: CHECKED	∅
ANSWER: Rejection	ANSWERED:	
Utterance 1 [neutral] Utterance 2 [assertive]	YES	∅ NO ∅
Utterance 3 [polite]		

3.2 Multimodal Interaction (II)

Apart from the automatically translated prepared utterances activated by the Users on both ends (2), the System also uses a Speech-to-Speech translation System with the use of the previously presented simplified form of Interlinguas, the S-ILTs (2), operating within a strict Directed Dialog framework. During the actual interaction, Users may speak, while at the same time they can activate the corresponding, text message (Table 4). Thus, the Interlingua processes spoken input whose content is identical, similar or even different than the generated text message, allowing both a controlled and a spontaneous type of information to be directed and evaluated by the Receiver. Additionally, it may be noted that written and spoken input may be compared to paralinguistic elements appearing on the screen.

The flow of the dialogs is controlled by the agenda in the template, indicating addressed and pending issues. The template-agenda contains the topics covered during the interaction and reminds Users, if a topic is not covered. A similar process is activated if opinions are requested. The template may be visible to the User, if requested (“open”, “close”). If an issue or a question is not addressed, the agenda on the template informs the User with a respective message, for example “Pending Issue: Terms and Conditions” or “No answer”. To repeat question press “R”. These messages

appear below the screen. In the case of an unaddressed issue, the System may remind the User for a predefined number of times after the response of the Receiver. Issues that are left unaddressed are saved in the agenda of the template. In the case of an unanswered question, especially, if opinions are asked, after a second attempt, the question is marked as unanswered in the agenda of the template.

Table 4. Chronological process of interaction

Before Interaction: Preparation/Editing Utterances
During Interaction: Prepared Utterances ->Text-to-Text Translation- Spontaneous Utterances ->Speech-to-Speech Translation
After Interaction: Template with points covered / Issues unaddressed / no answers Free input

3.3 Post-Interaction (III)

At the end of the interaction, closing remarks and any additional comments are processed as free input and saved in a wav.file to be processed after the interaction by a speech recognition (ASR) system and are subject to translation. After the interaction, the template also indicates unaddressed issues and unanswered questions, including unexpressed opinions.

4 Expression of Opinion

In multilingual HCI applications not limited to Task-Related Speech Acts and allowing the expression of views, opinions and spoken suggestions, as in the case of the present application, the relation of the semantic with the pragmatic level may create problems in some language pairs, since the “opinion” is expressed either directly or indirectly with significant differences between languages. Specifically, it may be noted that opinions in relation to a query or a specific issue may not always be expressed with a direct answer at a “Yes-No” question. Depending on the languages concerned, opinion may be contained within specific expressions or it may be avoided but indirectly expressed at the pragmatic level, for example, by avoiding to specifically address an issue.

In the proposed approach, the use of templates managed by the System allows the intervention of the System with inserted Non Task-Related Speech Acts (a) to control complex types of interaction concerning the expression of opinion in multiple languages and (b) to avoid the processing of additional Speech Acts containing culture-specific politeness features such as compliments and polite refusal in Chinese [12], [17] or highly-context-based politeness in Hindi [6]. Specifically, the template-agenda of the System contains fields to be filled in with specific types of opinion related to specific types of issues or queries. If these fields fail to be filled in, the System makes two

attempts to direct the Speaker to produce an answer. In the case in which opinions may be expressed with a direct answer at a “Yes-No” question or a “Wh-Question”, the rigid nature of the Directed Dialog structure allows the formulation of a Task-Related Speech Act (“Yes-No” question or “Request” Speech Acts). It may be additionally noted that the feasibility of the processing of this type of input by S-ILTs is directly related to the content of the Directed Dialogs, constraining User-input. For example, the System may ask: “What do you think? Please say “Yes” if you agree or “No” if you disagree” or “Please express your answer choosing one of the selected words from the template”. The “Yes” or “No” answer may be replaced by equivalent language-specific expressions in languages concerned.

Non Task-Related Speech Acts may be used by the System for the extraction of opinions, especially in the case in which a Task-Related Speech Act fails, for example, the “Reminder” or even the “Offer” Non Task-Related Speech Acts.

4.1 Indirect Expression of Opinion and Avoiding Expressing Opinion

In cases of languages processed in which opinions are frequently indirectly stated with the use of specific phrases or expressions, predefined expressions and phrases expressing opinions or views may be signaled and processed as additional keywords in the template-agenda. These expressions may include language-specific word-types related to connotative features such as adjectives and adverbs, verbs containing semantic features related to descriptive features, mode, malignant/benign action or emotional/ethical gravity, as well as some modal verbs [1]. These elements may be tagged as “non-neutral” opinion markers [4] in written text processed by the Text-to-Text Machine Translation System or in the proposed S-ILTs in the Speech-to-Speech Translation System. In respect to the proposed S-ILTs, expressions and phrases expressing opinions or views may be defined at a lexical level and processed by the Simple Interlinguas (S-ILTs). For example, in a Dialog System involving the expression of views, opinions and spoken suggestions [18], User actions are categorized as “propose” (a User proposes an idea with respect to a topic), “comment” (a User comments on a proposal, or answers a question), “acknowledgement” (a User confirms someone else’s comment or explanation, e.g., “yeah,” “uh huh,” and “OK”), “requestInfo” (a User requests unknown information about a topic), “askOpinion” (a User asks someone else’s opinion about a proposal), “posOpinion” (a User expresses a positive opinion, i.e., supports a proposal) and “negOpinion” (a User expresses a negative opinion, i.e., disagrees with a proposal)[18]. These elements may be registered in the System’s template-agenda as a positive, negative or other type of opinion.

In the multilingual interaction, there are cases in which the expression of opinion is completely avoided, for instance, the Users continue in not responding to a question, respond in an unexpected way or introduce a different topic. In this case, as described above, the template indicates unaddressed issues, unanswered questions and unexpressed opinions after the interaction for subsequent evaluation.

5 Conclusions and Further Research

For multilingual applications concerning business meetings and other complex types of interaction beyond purely Task-Related dialogs, the System is proposed to act as a mediator to control the dialog flow. The control of the interaction via the template-agenda is shared between the System, controlling interaction, and the Users, checking the issues addressed and activating prepared utterances.

The proposed framework allows the handling of opinion in routine tasks in business meetings to a relatively high extent, excluding the case of negotiations and business deals. However, it cannot cover all cases in which opinions or emotions are expressed, nor can it replace the benefits of an in person meeting with the parties concerned, with or without the presence of an interpreter. On the other hand, this framework also allows the evaluation of the communication immediately after the interaction. It is strictly sublanguage and domain-specific and requires some type of preparation from the Users, which may re-used for the same type of meetings in various languages with minor alterations or no alterations at all. The proposed approach allows reusability for standardized types of communication within a cross-linguistic and cross-cultural framework.

Further research is required in respect to the actual implementation of the proposed design in various languages as well as in relation to possible adaptation to other types of Service- Oriented Dialog Systems in different domains.

References

1. Alexandris, C.: English, German and the International “Semi-professional” Translator: A Morphological Approach to Implied Connotative Features. *Journal of Language and Translation* 11(2), 7–46 (2010)
2. Alexandris, C.: Speech Acts and Prosodic Modeling in Service-Oriented Dialog Systems. In: *Computer Science Research and Technology*. Nova Science Publishers, Hauppauge (2010)
3. Alexandris, C.: “Show and Tell”: Using Semantically Processable Prosodic Markers for Spatial Expressions in an HCI System for Consumer Complaints. In: Jacko, J.A. (ed.) *HCI 2007*. LNCS, vol. 4552, pp. 13–22. Springer, Heidelberg (2007)
4. Alexandris, C.: User Interface Design for the Interactive Use of Online Spoken German Journalistic Texts for the International Public. In: Stephanidis, C. (ed.) *Posters, Part I, HCII 2011*. CCIS, vol. 173, pp. 551–555. Springer, Heidelberg (2011)
5. D’Souza, R., Shivakumar, G., Swathi, D., Bhattacharyya, P.: Natural Language Generation from Semantic Net like Structures with application to Hindi. In: *Proceedings of STRANS- 2001: Symposium on Translation Support Systems*, Kanpur, India (2001)
6. Kumar, R.: A Politeness Recognition Tool for Hindi, with Special Emphasis on Online Texts. In: *Proceedings of the WWW PhD Symposium*, Hyderabad, India, March 28–April 1 (2011)
7. Levin, L., Gates, D., Lavie, A., Pianesi, F., Wallace, D., Watanabe, T., Woszczyna, M.: Evaluation of a Practical Interlingua for Task-Oriented Dialog. In: *Proceedings of ANLP/NAACL-2000 Workshop on Applied Interlinguas*, Seattle, WA (April 2000)

8. Malagardi, I., Alexandris, C.: Verb Processing in Spoken Commands for Household Security and Appliances. In: Stephanidis, C. (ed.) UAHCI 2009, Part II. LNCS, vol. 5615, pp. 92–99. Springer, Heidelberg (2009)
9. Matsumoto, N., Ueda, H., Yamazaki, T., Murai, H.: Life with a Robot Companion: Video Analysis of 16-Days of Interaction with a Home Robot in a “Ubiquitous Home” Environment. In: Jacko, J.A. (ed.) HCI International 2009, Part II. LNCS, vol. 5611, pp. 341–350. Springer, Heidelberg (2009)
10. Moeller, S.: Quality of Telephone-Based Spoken Dialog Systems. Springer, New York (2005)
11. Nottas, M., Alexandris, C., Tsopanoglou, A., Bakamidis, S.: A Hybrid Approach to Dialog Input in the CitizenShield Dialog System for Consumer Complaints. In: Proceedings of HCI 2007, Beijing, Peoples Republic of China (2007)
12. Pan, Y.: Politeness in Chinese Face-to-Face Interaction. Advances in Discourse Processes series V. 67. Ablex Publishing Corporation, Stamford (2000)
13. Schultz, T., Alexander, D., Black, A., Peterson, K., Suebvisai, S., Waibel, A.: A Thai speech translation system for medical dialogs. In: Proceedings of the conference on Human Language Technologies (HLT-NAACL), Boston, MA, USA (2004)
14. Wieggers, K.E.: Software Requirements. Microsoft Press, Redmond (2005)
15. Williams, J.D., Witt, S.M.: A Comparison of Dialog Strategies for Call Routing. International Journal of Speech Technology 7(1), 9–24 (2004)
16. Williams, J.D., Poupart, P., Young, S.: Partially Observable Markov Decision Processes with Continuous Observations for Dialog Management. In: Proceedings of the 6th SigDial Workshop on Discourse and Dialog, Lisbon (September 2005)
17. Yu, Z., Yu, Z., Aoyama, H., Ozeki, M., Nakamura, Y.: Capture, Recognition, and Visualization of Human Semantic Interactions in Meetings. In: Proceedings of PerCom., Mannheim, Germany (2010)
18. Zhu, H., Wei, L., Yuan, Q.: The sequential organization of gift offering and acceptance in Chinese. Journal of Pragmatics 32, 81–103 (2000)
19. The Agent-DYSL Project, <http://www.agent-dysl.eu/>
20. The HEARCOM Project, <http://hearcom.eu/main.html>
21. The ERMIS Project, <http://www.image.ntua.gr/ermis/>
22. The SOPRANO Project, <http://www.soprano-ip.org/>
23. The UNDL Foundation, The United Nations, <http://www.undlfoundation.org/undlfoundation/>

Linguistic Processing of Implied Information and Connotative Features in Multilingual HCI Applications

Christina Alexandris and Ioanna Malagardi

National University of Athens, Greece
calexandris@gs.uoa.gr, i.malagardi@gsrt.gr

Abstract. Implied information and connotative features may not always be easily detected or processed in multilingual Human-Computer Interaction Systems for the International Public, especially in applications related to the Service Sector. The proposed filter concerns the detection of implied information and connotative features in HCI applications processing online texts and may be compatible with Interlinguas including the signalization of connotative features, if necessary. The proposed approach combines features detected in the lexical and morpho-syntactic level, and in the prosodic and paralinguistic levels.

Keywords: Gricean Cooperativity Principle, online texts, Interlinguas, Morphology, prosodic and paralinguistic features.

1 Introduction

The management of implied information and connotative features concerns the Lexical-Semantic - Morphosyntactic Level as well as the Prosodic Level and the Paralinguistic Level. For the management of these features, three types of strategies may be differentiated, depending on the type of application and task-type: excluding, detecting or integrating implied information and connotative features. The exclusion of implied information and connotative features is typically applied in task-related monolingual or multilingual applications in Human Computer Interaction (HCI) Systems, including monolingual Dialog Systems and HCI Systems for the International Public based on Machine Translation (MT). Such applications often involve Controlled Languages and/or a Directed Dialog and System-Initiative-based [12] approach.

Furthermore, some types of monolingual Dialog Systems or other monolingual Human Computer Interaction Systems may pose problems for the International Public, especially in the service sector, as Human-Computer interactions go global [6] and the alternative practice of monolingual variations in the respective languages of a standardized HCI System is proposed. This practice often involves the integration of the language and culture-specific implied information and connotative features.

However, implied information and connotative features may not always be easily avoided in multilingual Human-Computer Interaction Systems for the International Public, especially in applications related to the service sector. In some cases, it may be necessary for such HCI Systems using Machine Translation to include the strategy of detecting implied information and connotative features, according to the type and purpose of the application.

2 Application Types

The proposed filter may be compatible with the processing of online texts for the detection of implied information and connotative features in interactive HCI applications processing monolingual online texts for the International Public. This is especially important for professionals, such as journalists, economists and other professionals working with multilingual written and spoken texts, either in specialized domains or in general fields such as business meetings, business transactions and online financial news. It should be noted that International Users may be very fluent in a foreign language or more than one foreign languages, but are either non-native speakers and/or often lack the necessary exposure to the culture related to a foreign language concerned to easily perceive all types of implied information and connotative features.

In online texts, the word groups with implied information and connotative features may be automatically signalized by an online tagger in the proposed filter and in some cases, they be also be provided in combination with the proposed filter as a set of guidelines to International Users, activated according to request.

The proposed filter may also be compatible with Interlinguas. Traditional Interlinguas are geared towards the exclusion and of connotative features [11] [15]. However, in the present approach connotative features may be signalized in Interlinguas, if necessary.

Interlinguas may allow extra tags at a lexical level to be placed with connotative features that may correspond to word groups presented. The connotative features may be automatically signalized operating on detection at morpheme-level or word-level, based on interaction with a database. In addition to detecting implied information and connotative features, the signalization of language-specific prosodic emphasis and other paralinguistic elements at word-level in an Interlingua may play a significant role in the information content of spoken utterances, possibly also in Tonal Languages such as Thai or Chinese.

For all applications concerned, the language-specific “filter” is proposed for the detection or for the integration of implied information and connotative features which may be adapted to the needs of the HCI applications concerned. The filter may be activated or deactivated when processing online texts or during spoken interaction with Interlinguas.

Specifically, in the approach presented, the database and respective tag-set with the word groups and related elements concerned may be used in a variety of multilingual applications, in particular, in the interactive processing of online written and spoken texts, as well as in the processing of Interlinguas and/or its integration in monolingual variations of dialogs. The proposed database and tag-set “filter” is based on features from English, German and Modern Greek, however, it can be adapted and extended to other languages and language groups.

Table 1. Interaction with database and tag-set “filter”

Input Type (During Interaction):	Filter:
INTERLINGUAS (ILT OR S-ILT) Speech-to-Speech Translation	Activate – Deactivate ⊗ [open/close]
ONLINE TEXT	

3 Filter for Linguistic Processing

The proposed language-specific “filter” constitutes a simple and extendable database constructed on ontological and pragmatic principles [1] [3] and respective tag-set with basic types of word groups related to implied information and connotative features. In particular, the connotative features, implied information and other types of elements may be automatically signalized at morpheme-level or word-level.

The proposed filter is composed of three tiers, the Lexical Tier, the Prosodic Tier and the Paralinguistic Tier and allows the combination of all tiers and respective linguistic and paralinguistic elements to evaluate implied information and connotative features. The combination of all tiers related to the respective Lexical-Semantic - Morphosyntactic Level (Lexical Tier), the Prosodic Level (Prosodic Tier) and the Paralinguistic Level (Paralinguistic Tier) allows a significant extent of coverage of implied information and connotative features for the International Public.

We note that the intensity of the connotative features may be stronger if detected in all three levels, for example, if a word containing (lexical-semantic) connotative features receives prosodic emphasis or is accompanied by additional connotative information in the Paralinguistic Level.

3.1 Tags Related to the Paralinguistic Tier

The Paralinguistic Tier combines paralinguistic elements with spoken words. Paralinguistic Elements are usually language-specific and may vary across cultures. Typical examples of paralinguistic elements are facial expressions such as the raising of eyebrows and frowns, as well as body movements such as gestures related to the hands or nodding of head. Tags for paralinguistic elements may be the annotations “[raising eyebrows]”, “[frown]” and “[nod]”. Additional features may be added, for example, in respect to speed, such as the annotation “[nod-quick]”, in addition to highly language (and culture) specific variations.

3.2 Tags Related to the Prosodic Tier

In the Prosodic Tier, the proposed filter as an annotation module combines spoken words with prosodic elements, such as prosodic emphasis signalizing stressed

elements or a casual attitude: “Stress”, “Casual” [1]. For spoken Machine Translation, types of Interlinguas (ILTs) allowing the recognition and isolated processing of keywords at a lexical level, such as Simple Interlinguas [13], facilitate the signalization of prosodic features detected at word level, such as prosodic emphasis. For example, the marker prosodic emphasis [+STRESS] can be used as an additional paralinguistic marker on the lexical element of the Interlingua or in the online spoken text. The marker [+CASUAL] may be used for a casual tone, whereas the [+NON-NEUTRAL] is used on other types of prosodic elements related to implied connotative features on the Prosodic Level.

3.3 Tags Related to the Lexical Tier and Pragmatic Principles

In the Lexical Tier, corresponding to tags concerning the Lexical Level, the word groups with connotative features are differentiated according to criteria related to Pragmatics, namely the flouting of the Maxims stated in the Gricean Cooperativity Principle [8] [9].

In respect to the Gricean Maxim of Quality, namely “Do not say what you believe to be false” and “Do not say that for which you lack adequate evidence”, in the case in which the Maxim is not flouted for the purpose of propaganda, it is observed that information presented in a form characterized as flouting the Gricean Maxim of Quality often contains superfluous elements flouting the Gricean Maxim of Quantity. This relationship applies in a similar way to information flouting the Gricean Maxim of Manner and, specifically, “Avoid obscurity of expression”, “Avoid ambiguity” and “Be orderly”, where, unless propaganda or similar communicative targets are involved, superfluous information (flouting the Gricean Maxim of Quantity) is often encountered in written and spoken texts where there is flouting of the Gricean Maxim of Manner. The Gricean Maxim of Manner also includes the part “Be brief” (avoid unnecessary prolixity) which partially coincides with the Gricean Maxim of Quantity and is thus directly related to the avoidance of superfluous information.

Specifically, these word groups to be detected or integrated in multilingual applications concern both specific types of semantic features related to the superfluous information connected to the above-described Maxims, such as mode, malignant/benign action or emotional/ethical gravity, as well as particular types of grammatical features in verbs, adjectives, adverbials and in specific types of suffixes and particles. The language-specific tag set of word groups ranges from the more evident yet less frequent strong or emotional expressions to the less obvious and commonly-occurring word categories constituting word groups related to implied information and connotative features.

These word categories function as subtle hints or tell-tale signs of connotative elements and implied information and may sometimes be especially problematic to the International Public.

In other words, the overall context of the written and spoken text may be described as containing a subset of word-types, coinciding with superfluous information in the text and indicating emotionally and socio-culturally “marked” elements constituting implied information and connotative features and expressing style and overall spirit of

the author, speaker and the intended readership or audience. Thus, the criteria for determining tagged word types with implied and connotative features are related to Pragmatics, in particular, the flouting of the Gricean Cooperativity Principle [8] [9].

Table 2. Tiers of the proposed database and tag-set “filter”

Features:

Paralinguistic Tier
(Paralinguistic Level)
 [raising eyebrows] [frown] [nod-quick]

Prosodic Tier
(Prosodic Level)
 [±STRESS] [±CASUAL] [±NON-NEUTRAL]
 (other types of prosodic elements)

Lexical Tier
(Lexical Level/Word Level – Morphological Level)
 [sem-expl-conn]
 [sem-impl-conn]
 [prag-conn:±emph]
 [prag-conn:subj]
 [prag-conn:modl]

4 Connotative Feature Types

Tags corresponding to word groups with implied information or connotative features in their semantic content may be divided into word categories where connotative features are detected at a word level (Lexical Tier) and word categories where connotative features are detected in a morphological level (morpheme level) (Lexical Tier). Connotative features detected at word level or at the morphological level are either related to word groups whose semantic content is related to connotatively emotionally, and socio-culturally “marked” elements (Semantic Content categories) or word groups whose pragmatic usage concerns connotative features and implied information (Pragmatic Usage categories). From the aspect of Prosody, it is observed that the word categories detected at a word level are sensitive to prosodic emphasis. In contrary, the word categories detected at the morphological level are not affected by prosodic emphasis.

4.1 Word Level and Semantic Content

Word categories with connotative features detected at the word level (Lexical Tier) and whose semantic content is related to connotatively emotionally, and socio-culturally “marked” elements constitute word groups with evident or explicit connotative features.

Typical examples of word groups with explicit connotative features are the grammatical categories of adjectives and adverbials, containing semantic features related to (i) descriptive features (ii) mode (iii) malignant/benign action or (iv) emotional/ethical gravity [3].

The evident connotative features of the above-described word categories may be formalized as special features linked to the respective word category type. Specifically, the feature [sem-expl-conn] may be appended to these categories and matched on a word-level, being directly matched to the entire word. In respect to the Prosodic Level of this group of word categories, prosodic emphasis may emphasize or intensify the semantics of the emphasized word without determining the semantic content. The word groups with evident connotative features may be classified as “Prosodically Sensitive” words.

Table 3. Connotative Feature Types

		CONNOTATIVE	
		DETECTION:	FEATURE TYPE: PROSODY:
WORD CATEGORY	→	WORD LEVEL →	SEMANTIC CONTENT “Prosodically Sensitive”
		→	PRAGMATIC USAGE “Prosodically Sensitive”
	→	MORPHO- LOGICAL LEVEL →	SEMANTIC CONTENT “Prosodically Independent”
		→	PRAGMATIC USAGE “Prosodically Independent”

4.2 Morphological Level and Semantic Content

Word categories whose connotative features are detected in the morphological level (Lexical Tier) and whose semantic content is related to connotatively emotionally and socio-culturally “marked” elements also may be referred to as word groups with implicit connotative features [2].

Word groups with implied connotative features include the grammatical categories of verb-stems (or nominalizations of verbs) containing semantic features (including implied connotations in language use) related to (i) mode (ii) malignant/benign action or (iii) emotional/ethical gravity, as well as nouns with suffixes producing diminutives, derivational suffixes resulting to a (ii) verbalization, (iii) an adjectivization or (iii) an additional nominalization of proper nouns [1] [3].

The implicit connotative features of the above-described word categories may be formalized as special features linked to the respective word category type. Specifically, the feature [sem-impl-conn] may be appended to these categories, matched on a morphological level, on the verb-stem of verbs containing semantic features related to mode, malignant/benign action or emotional/ethical gravity, and on the suffix of nouns with suffixes producing diminutives [2]. Additionally, for the appending of the [sem-impl-conn] feature, verb-stems are compared with derivational suffixes resulting to a nominalization of verbs (excluding derivational suffixes producing participles and actor thematic roles). Stems of proper nouns are compared with derivational suffixes resulting to a verbalization or adjectivization of proper nouns [3].

In respect to the Prosodic Level, the presence or absence of prosodic emphasis on words of this word group only effects the semantic interpretation of the entire phrase or sentence in which they belong. A significant percentage of these words are nouns or verbs and they may constitute sublanguage-specific keywords. Prosodic emphasis on keywords focuses on the basic content of the utterance, for example, whether it is an action in question, in the case of a verb, or a specific object in question, in the case of a noun. Prosodic emphasis on the word elements of this category, which may be classified as “Prosodically Independent”, is sentence dependent and highly sublanguage- and application-specific.

4.3 Word Level and Pragmatic Usage: Adverbials and Particles

Word categories with connotative features detected at word level (Lexical Tier) and whose pragmatic use concerns connotative features and implied information involve language-specific sets of adverbials, discourse particles or other language-specific grammatical categories.

The feature [prag-conn:±emph] is matched on a word-level being directly matched to the adverbial or particle used in languages such as English (“so”) or German (“eben”, “gleich”) for an emphatic or casual/spontaneous effect or the discourse particle identified as a “politeness marker” in Modern Greek (“Πείτε μου”).

In spoken language, these adverbials and particles may either be used to emphasize the semantic content of the spoken phrase or sentence (emphasis [+emph]), to allow a more casual or spontaneous effect of the overall spoken utterance (casual, [-emph]) or to achieve politeness (politeness-markers, [-emph]).

Regarding the Prosodic Level, for discourse particles identified as “politeness markers”, the absence of prosodic emphasis signals them as politeness markers, while with the presence of prosodic emphasis they only have the property of discourse particles [4][5]. Similarly, for adverbials and particles in languages such as English

(“so”) or German (“gleich”), the absence of prosodic emphasis signalizes a casual or spontaneous effect, which is not achieved with the presence of prosodic emphasis [3].

Previous studies have demonstrated a differentiation between specific word categories in which prosodic emphasis does not determine their semantic content and word categories whose semantic content may be determined by prosodic emphasis [5]. In the present case involving adverbials and particles, the semantic content is not entirely determined by the presence or absence of prosodic emphasis, however, the pragmatic features within the utterance and the related connotative aspects are affected. The group of word categories whose semantic content may be affected by prosodic emphasis is classified as “Prosodically Sensitive” words [5].

4.4 Morphological Level and Pragmatic Usage: Other Grammatical Features

Word categories with connotative elements detected at the morphological level (Lexical Tier) and whose pragmatic use concerns connotative elements and implied information concern various types of grammatical features. Grammatical features inherently present in languages may contain implied semantic and connotative information which is not always easily detected or successfully managed in the translation process.

Examples of ambiguity related to implied or hypothetical actions are modal verbs and verbal adjectives. Apart from their literal meaning to express a suggestion or a prediction, modal verbs in English and German such as “should” (“soll” in German) or “would” (“wuerde” in German) are often used as understatements, an implied intention, sometimes even irony. These grammatical categories may be in many cases especially problematic to the International Public, both in written language and in spoken language.

Another example of inherent grammatical features and implied information and connotative features are the suffixes in specific verb groups of pro-drop languages. The connotative feature of politeness or friendliness can be expressed in the form of a relationship between subject and object is detected in Greek verb suffixes.

Specifically, we note that, in Greek, as a verb-framed and pro-drop language (like Spanish or Italian), the verbal features in verb’s suffix imply the subject. This morphological characteristic affects the semantics and connotative features of certain verb categories (verb stems), especially verbs expressing a service or any benign action concerning an object or the verb’s subject or both the subject and object of the action expressed. With this way, a relation between the subject and the object is expressed, signaling politeness, especially in spoken language, if receiving prosodic stress [13]. Emphasis is placed on the User’s wish or response. For example, the Greek verbs “theleis” (“[you] want”) and “olokli’rosate” (“[you] have finished-completed” [your input]) is equivalent to the verbs “want” and “finished” in English respectively.

Features appended to these categories, are the feature [prag-conn:subj] is matched on a Morphological Level, being matched on the suffix of verbs of verb-framed and pro-drop languages.

For modal verbs, the subset of modal verbs containing likely connotative features is signalized by the feature [prag-conn:modl] [3]. Feature types allow an automatic

grouping of verb groups and other word groups [10] [14] and may also be retrieved with the help of Wordnets. Additional feature types may be added, according to the language concerned.

In respect to the Prosodic Level, words of this word group are unaffected by the presence or absence of prosodic emphasis in respect to their semantic content, constituting a “Prosodically Independent” category.

As a final comment in respect to the observed “Prosodically Independent” categories related to Semantic Content and to Pragmatic Usage, the Morphological Level appears to be opaque to any prosodic interference affecting semantic content. However, further research is required to evaluate this observation.

5 Conclusions and Further Research

The proposed approach involves a database constructed on ontological and pragmatic principles and respective tag-set with basic types of word groups related to implied information and connotative features, constituting a language-specific “filter” for HCI applications.

The proposed language-specific “filter” may operate simultaneously on the Morphosyntactic and Lexical-Semantic Level, the Prosodic Level and the Paralinguistic Level. Additionally, the proposed approach allows the processing of implied information and connotative features related in the combination of multiple linguistic levels, enabling access to complex types of implied information and connotative features.

Furthermore, with the proposed approach is targeted to allow flexibility in respect to various languages, serving as an onset for a cross-linguistic approach. In particular, the language-specific annotation containing implied information and connotative features may be inserted and signaled at the Lexical Tier and the Prosodic Tier, as well as the Paralinguistic Tier, providing a flexible framework for the processing of various languages other than the languages presented.

It additionally may be noted that the management of implied information and connotative features in word groups may contribute to ambiguity resolution in semantic webs and in some cases even in the social semantic web, especially in respect to tags [7].

An additional target is to gain insight for the formalization of a basic framework for processing similar or contrasting linguistic and cultural features of other language families. Further research including a comparison with other languages and language families may allow the integration of additional features and/or aspects in the proposed general framework.

References

1. Alexandris, C.: User Interface Design for the Interactive Use of Online Spoken German Journalistic Texts for the International Public. In: Stephanidis, C. (ed.) Posters, Part I, HCI 2011. CCIS, vol. 173, pp. 551–555. Springer, Heidelberg (2011)
2. Alexandris, C.: English, German and the International “Semi-professional” Translator: A Morphological Approach to Implied Connotative Features. *Journal of Language and Translation* 11(2), 7–46 (2010)

3. Alexandris, C.: *Speech Acts and Prosodic Modeling in Service-Oriented Dialog Systems*. In: *Computer Science Research and Technology*. Nova Science Publishers, Hauppauge (2010)
4. Alexandris, C.: "Show and Tell": Using Semantically Processable Prosodic Markers for Spatial Expressions in an HCI System for Consumer Complaints. In: Jacko, J.A. (ed.) *HCI 2007*. LNCS, vol. 4552, pp. 13–22. Springer, Heidelberg (2007)
5. Alexandris, C.: *Word Category and Prosodic Emphasis in Dialog Modules of Speech Technology Applications*. In: Botinis, A. (ed.) *Proceedings of the 2nd ISCA Workshop on Experimental Linguistics, ExLing 2008*, Athens, Greece (2008)
6. Altman-Klein, H., Lippa, K., Lin, M.H.: *As Human-Computer Interactions Go Global*. In: Hayes, C.C., Miller, C.A. (eds.) *Human-Computer Etiquette*, Taylor and Francis, Boca Raton, FL (2011)
7. Breslin, J.G., Passant, A., Decker, S.: *The Social Semantic Web*. Springer, New York (2009)
8. Grice, H.P.: *Logic and conversation*. In: Cole, P., Morgan, J. (eds.) *Syntax and Semantics*, vol. 3. Academic Press, New York (1975)
9. Hatim, B.: *Communication Across Cultures: Translation Theory and Contrastive Text Linguistics*. University of Exeter, Exeter (1997)
10. Kontos, J., Malagardi, I., Pegou, M.: *Processing of Verb Definitions from Dictionaries*. In: *Proceedings of the 3rd International Conference in Greek Linguistics*, Athens, pp. 954–961 (1997) (in Greek)
11. Levin, L., Gates, D., Lavie, A., Pianesi, F., Wallace, D., Watanabe, T., Woszczyna, M.: *Evaluation of a Practical Interlingua for Task-Oriented Dialog*. In: *Proceedings of ANLP/NAACL-2000 Workshop on Applied Interlinguas*, Seattle, WA (April 2000)
12. Jurafsky, D., Martin, J.: *Speech and Language Processing, an Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, 2nd edn. Prentice Hall series in Artificial Intelligence. Pearson Education, Upper Saddle River (2008)
13. Malagardi, I., Alexandris, C.: *Verb Processing in Spoken Commands for Household Security and Appliances*. In: Stephanidis, C. (ed.) *UAHCI 2009, Part II*. LNCS, vol. 5615, pp. 92–99. Springer, Heidelberg (2009)
14. Malagardi, I., Kontos, J.: *Motion Verbs and Vision*. In: *Proceedings of the 8th Hellenic European Research on Computer Mathematics & its Applications Conference (HERCMA 2007)*, Athens (2007), <http://www.aueb.gr/pympe/hercma/proceedings2007>
15. Schultz, T., Alexander, D., Black, A., Peterson, K., Suebvisai, S., Waibel, A.: *A Thai speech translation system for medical dialogs*. In: *Proceedings of the Conference on Human Language Technologies (HLT-NAACL)*, Boston, MA, USA (2004)

Investigating the Impact of Combining Speech and Earcons to Communicate Information in E-government Interfaces

Dimitrios Rigas¹ and Badr Almutairi²

¹ INSPIRE, University of West London, London W5 5RF, UK
Dimitrios.Rigas@uwl.ac.uk

² Faculty of Technology, De Montfort University, Leicester LE1 9BH, UK
Badr@dmu.ac.uk

Abstract. This research investigates the use of multimodal metaphors to communicate information in the interface of an e-government application in order to reduce complexity in the visual communication by incorporating auditory stimuli. These issues are often neglected in the interfaces of e-government applications. This paper investigates the possibility of using multimodal metaphors to enhance the usability and increase the trust between the user and the application using an empirical comparative study. The multimodal metaphors investigated include text, earcons and recorded speech. More specifically, this experiment aims to investigate the usability in terms of efficiency, effectiveness and user satisfaction in the context of a multimodal e-government interface, as opposed to a typical text with graphics based interface. This investigation was evaluated by 30 users and comprised two different interface versions in each experimental e-government tool. The obtained results demonstrated the usefulness of the tested metaphors to enhance e-government usability and to enable users to attain better communicating performance. In addition empirically derived guidelines showed that the use of multimodal metaphors in an e-government system could significantly contribute to enhance the usability and increase trust between a user and an e-government interface. These results provide a paradigm of a design framework for the use of multimodal metaphors in e-government interfaces.

Keywords: e-government, Recorded Speech, Earcons, Multimodal, Trust, HC1.

1 Introduction

For the time being, most of web interfaces applications are visually crowded and difficult to communicate the intended message correctly to users via the visual channel. Therefore, other human senses could be involved in human computer interaction to employ more interaction metaphors within the visual channel, the auditory channel or both. This research describes an empirical exploration that has been carried out to investigate the usability aspects of an e-government interface that incorporates a combination of typical text with multimodal metaphors such as recorded speech. The main question asked in this study is whether the inclusion of these metaphors can

enhance usability and communication with the user. A secondary question relates to the contributing role that each of these multimodal metaphors can play in the expected enhancement. An e-government experimental platform, with two interface versions, was developed to serve as a basis for this investigation. The e-government software solution described uses an input interface to send messages and an output interface to receive messages. The study involved two groups of users (one group for each interface version) in which the usability performance of the two groups, in terms of efficiency, effectiveness, and user satisfaction was compared.

1.1 Literature Review:

Usability Evaluation in e-government Interfaces

Usability is one of the most important factors to evaluate Human-Computer Interaction [1] and software quality [2]. It can be defined as the “*extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction*” [3]. The effective system can be implemented and developed only by understanding better government websites, expectations of the users' under the citizen-centric approach, and the barriers that might hinder these interfaces in order to provide the desired services through the Internet. This technology can be used to improve the efficiency of governments; improve interaction between government and public, facilitating economic development, reduce costs, and move towards meeting citizens' expectations for service delivery, by facilitating the process of administrative procedures [4, 5].

Multimodal Interaction

Rigas and Memery stated that “*the auditory channel, as a whole, has been neglected in the development of user-interfaces, possibly because there is very little known about how humans understand and process auditory stimuli*” [6]. Interfaces that offer interaction using more than one channels of communication are often more usable. Rigas et al, suggest that the use of multimodal metaphors in application interfaces increases usability and the volume of information that can be communicated to the user [7, 8, 9]. Also, they found that the use of speech and other auditory stimuli in an interface helped users to make fewer mistakes and reduced the time taken when accomplishing their tasks [10]. Several other studies have been carried out to test the use of multimodal metaphors in visual user interface and to evaluate and examine the affect of these metaphors on the usability of computer applications [11, 12].

2 Multi-Modal E-government Experimental Platform (MMEGP)

The main aim of this experiment was to measure the impact of combining recorded natural speech and earcons on the usability of e-government interfaces. It also aimed to evaluate the extent to which the addition of these multimodal metaphors can affect

the ability to communicate with users. More specifically, this experiment is aimed at evaluating the efficiency, effectiveness and user satisfaction of a multimodal e-government interface, as opposed to a typical text based interface. An additional aim was to explore these usability factors with different tasks in terms of complexity (i.e. easy, moderate and difficult) and message types (suggestions, complaints and comments) using both input and output and question types (i.e. recall and recognition). Therefore, this experiment aimed to investigate usability aspects as well as communication performance of e-government interfaces that combine text, recorded speech and earcons in order to also improve trust between users and the e-government interface.

Fig. 1. Multi-Modal e-government Experimental Platform (MMEGP) showing the input interface

Given the aims, the objectives of this study were to measure the performance of the users in terms of efficiency (time taken by users to complete tasks), effectiveness (successfully completed tasks by users), and user satisfaction by requesting users to rate the communication metaphors used in the platform.

Figure 1 shows an example screenshot of the multimodal e-government interface. Creation of the involved multimodal metaphors was primarily based on the connection between these interaction metaphors and the information being delivered. This connection also considered the previous interface that demonstrated the usefulness of multimodal interaction. The e-government interface contained information which was delivered in a textual way with recorded speech and earcons. Information could be communicated by the visual channel and by making use other communication channels in the interaction process (e.g. recorded speech, earcons and images). Guidelines for multimodal information presentation [13] and multimodal user interface design were followed. For example, the multimodal input and output was used to widen the bandwidth of information transfer [13, 14]. Also, graphical displays, speech messages

were combined to obtain an effective presentation [15] in a way that speech can be used to transmit short messages.

2.1 Variables

The variables considered in the experimental design can be classified into three types which are: independent variables, dependent variables and controlled variables.

Table 1. Independent variables considered in the experiment

Variable Code	Variable	Levels	State 1	State 2	State 3
IV 1	Presentation method	2	TOEGP	MMEGP	
IV 2	Message complexity	3	Easy	Moderate	Difficult
IV 3	Message type	3	Suggest	Complain	Comment

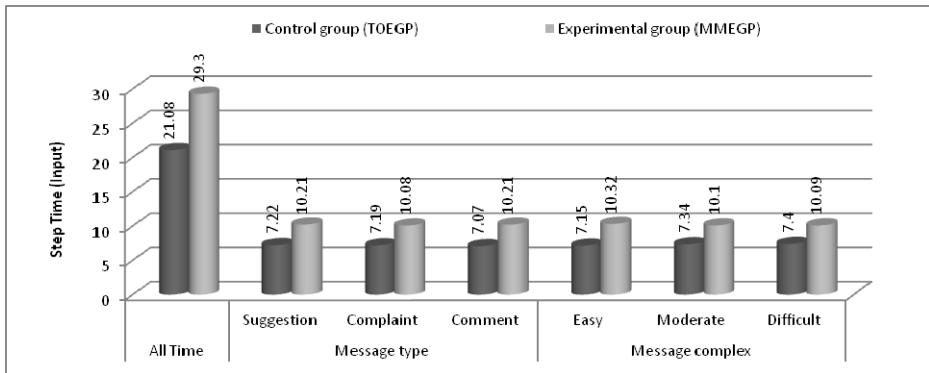


Fig. 2. Mean values of time taken by users in both groups to enter all tasks, grouped by message complexity and message type for the (input interface)

2.2 Efficiency

The time spent to enter message tasks and answer the required questions was used as a measure of efficiency. This measure was considered for all tasks for the input interface and for the output interface (according to the question type, recall and recognition), message complexity, as well as for each task and for each of the users in both groups.

A figure 2, 3 shows the time taken by users to complete the various types of tasks. It can be seen that the use of recorded speech was more efficient, as tasks took less time - unlike other groups which took more time to read the tasks.

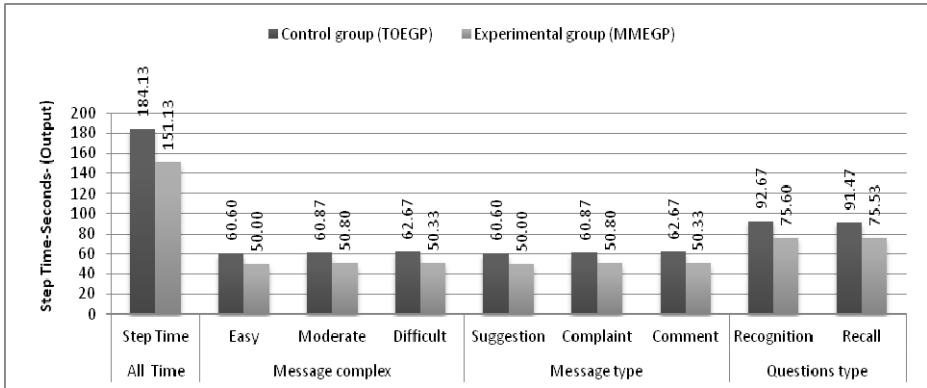


Fig. 3. Mean values of time taken by users in both groups to enter all tasks, grouped by message complexity and message type for the (output interface)

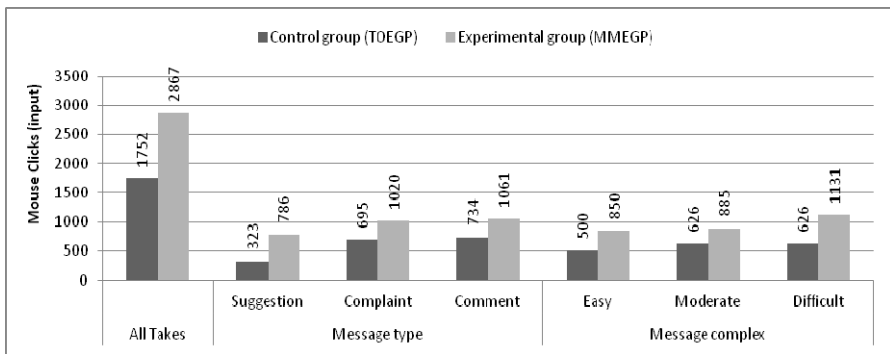


Fig. 4. Percentage of mouse clicks performed by users in both groups

2.3 Effectiveness

The numbers of correctly completed tasks were used to measure the effectiveness. This measure was considered for all messages and all the questions, according to the question type (recall and recognition) and message complexity (easy, moderate and difficult) and message type (suggestion, complain and comment), as well as for each user in both control and experimental groups.

Figure 4 shows the percentage of mouse clicks to enter messages for all tasks for the TOEGP and MMEGP. It can be noted that users of the TOEGP used less mouse clicks than users of the MMEGP. This was due to the requirement when using the input interface to enter text only, in contrast to the experimental group that required users to enter text and spoken speech. Figure 4 shows that the users of the TOEGP performed better than the users of the MMEGP with regard to the number of mouse clicks for all messages. The mean number of mouse clicks for the MMEGP was (2867) more than that attained in the TOEGP (1752) for all messages. The t-test results showed that the difference in mouse clicks between MMEGP and TOEGP was significant ($t(16)$, $MD=-2.9$, $p<0.05$). As a result, the MMEGP users outperformed the users of the TOEGP, who send their information using text only.

It can be seen that users of the TOEGP exceeded MMEGP users in terms of the number of mouse clicks used to enter messages for all tasks. The multimodal metaphors applied in the MMEGP assisted in reducing the number of mouse clicks used for the required tasks in the input interface.

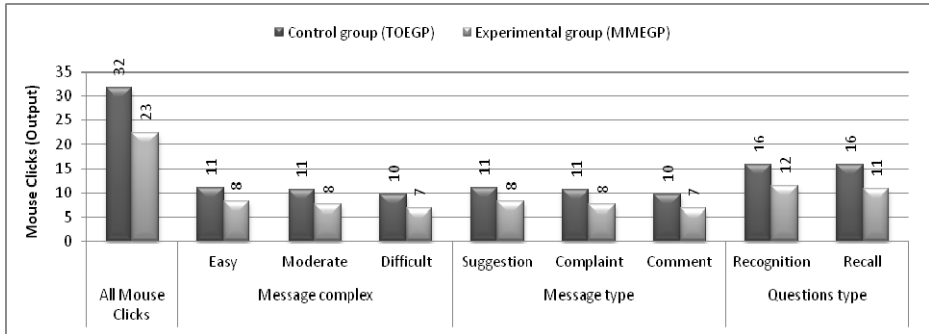


Fig. 5. The mean number of mouse clicks by users in both groups to enter message for all the tasks

Figure 5 shows that users of MMEGP performed better than the users for TOEGP in terms of the number of mouse clicks used for all messages. The mean number of mouse clicks used in the MMEGP (23) was less than that used in the TOEGP (32) for all messages in the output interface. The t-test results showed that the difference in mouse clicks between MMEGP and TOEGP was significant ($t(6)$, $MD=9$, $p<.05$). As a result, TOEGP outperformed the users of the MMEGP when received the messaging information via text with metaphors. The correct combination of more than one communication metaphor of different channels in the MMEGP helped users in the experimental group to discriminate between the different types of information delivered by each of the recorded speech extracts, thus enabling them to understand this information in a short time period and reduced the number of mouse clicks. In summary, the multimodal interaction metaphors used in the MMEGP were more effective in communicating and considerably assisted the users in the experimental group to achieve a higher effectiveness rate, as opposed to the control group users using the output interface.

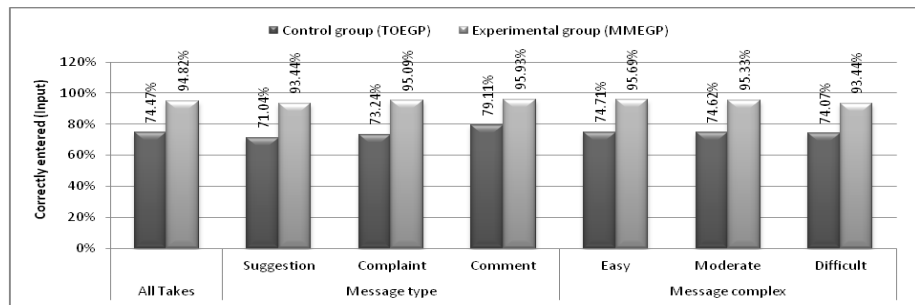


Fig. 6. Percentages of correctly completed tasks by type

By analysing the “correctly entered” measure we can find the percentage of users that entered the correct message at the input for all tasks. Figure 6 shows the percentage of test results of information correctly entered for all tasks in the TOEGP and MMEGP.

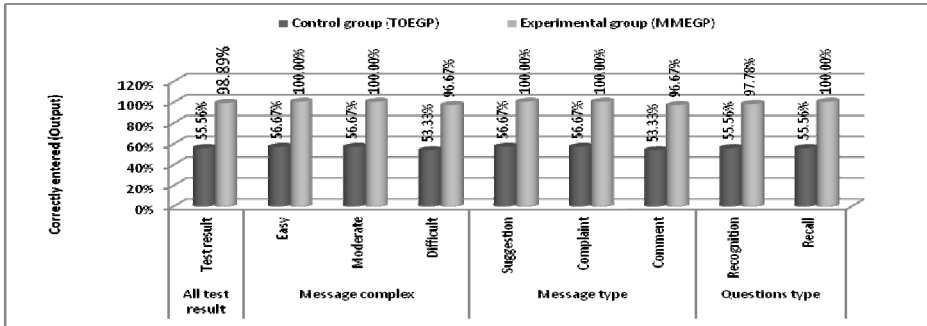


Fig. 7. Percentage of correctly completed tasks by correctly entered for users in both groups for the output interface

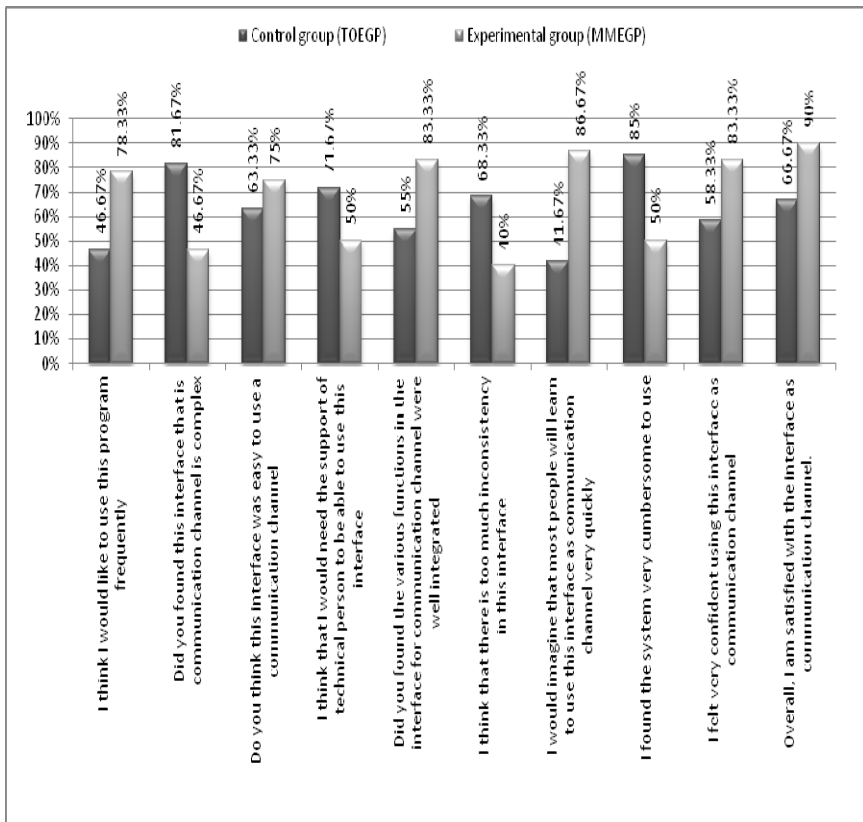


Fig. 8. Percentages of users agreeing to each statement of satisfaction for both TOEGP and MMEGP groups

Figure 7 shows that users of the MMEGP completed more tasks successfully than the TOEGP users, in terms of the number of correctly entered messages for tasks using the output interface. The MMEGP was more effective in communicating and considerably assisted the users in the experimental group to achieve a higher effectiveness rate, as opposed to users in the control group.

2.4 User Satisfaction

The user satisfaction with regard to the different aspects of the applied e-government platform was measured for both groups in terms of users' answers to the post-experimental questionnaire which consisted of 10 statements related to the ease of use.

The overall satisfaction score for each user was calculated using the SUS (System Usability Scale) method. The mean satisfaction score for the users in the experimental group was 68.33%, compared to 63.83% for the users in the control group. In other words, the MMEGP was more satisfactory for users than the TOEGP.

3 Concluding Summary

This paper examined the impact of multimodal interaction metaphors for ease of use in terms of efficiency, effectiveness and user satisfaction and the communication performance of an e-government experimental interface. This study has been implemented by developing two different versions of the experimental e-government platform. The results obtained from this experiment confirm that multimodal metaphors do in fact help to improve the usability of e-government interfaces, and reduce the time needed for users to respond to messages, and allow users to undertake activities more accurately, and make use of the interface more pleasing and satisfactory.

We can therefore conclude that the multimodal metaphors tested can contribute greatly to improving the performance of users' communication and ease of use of e-government interfaces in terms of effectiveness, efficiency and user satisfaction. It is therefore proposed to include multimodal metaphors in e-government interfaces and this need to be taken in mind when designing such interfaces. This e-government interface approach is gaining popularity among the providers of e-government services. Its importance from the users' point of view has become the main concern for e-government services.

References

1. Nielsen, J.: Usability Engineering. Academic Press Inc., US (1993)
2. Costabile, M.F.: Usability in the Software Life Cycle. Handbook of software engineering and knowledge engineering 1, 179–192 (2001)
3. ISO, ISO 9241: Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs) - Part 11: Guidelines on usability, p. 2, (1998)
4. Alzahrani, A.: Web-based e-Government services acceptance for G2C, A structural equation modeling approach. De Montfort University (2011)

5. Alsaghier, H., Ford, M., Nguyen, A., Hexel, R.: " Conceptualising Citizen's Trust in e-government: Application of Q Methodology". *Electronic Journal of E-Government* 7(4), 295–310 (2009)
6. Rigas, D., Memery, D., et al.: Experiments in using structured musical sound, synthesised speech and environmental stimuli to communicate information: is there a case for integration and synergy? *Intelligent Multimedia, Multimedia, Video and Speech processing* (2001)
7. Rigas, D.: *Guidelines for Auditory Interface Design: An Empirical Investigation*. PhD thesis, Loughborough University of Technology (1996)
8. Rigas, D., Alty, J.L.: Using sound to communicate program execution. In: *Proceedings of the 24th EUROMICRO Conference*, vol. 2, pp. 625–632 (1998)
9. Rigas, D., Hopwood, D.: The Role of Multimedia in Interfaces for On-Line Learning. In: *9th Panhellenic Conference on Informatics (PCI 2003)*, Thessaloniki, Greece (2003)
10. Rigas, D.: *Guidelines for Auditory Interface Design: An Empirical Investigation*. PhD thesis, Loughborough University of Technology (1996)
11. Michaelis, P.R., Wiggins, R.H.: A human factors engineer's introduction to speech synthesizers. *Directions in Human Computer Interaction* (1982)
12. Alotaibi, M., Rigas, D.: A Usability Evaluation of Multimodal Metaphors for Customer Knowledge Management. *International Journal of Computers and Communications* 2 (2008)
13. Oviatt, S.: *Multimodal Interfaces, The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications* (2003)
14. Sarter, N.B.: Multimodal information presentation: Design guidance and research challenges. *International Journal of Industrial Ergonomics* 36, 439–445 (2006)
15. Bonebright, T.L., Nees, M.A., Connerley, T.T., McCain, G.R.: Testing the effectiveness of sonified graphs for education: A programmatic research project. In: *International Conference on Auditory Display*, Espoo, Finland (2001)
16. Ciuffreda, A.: An Empirical Investigation in Using Multi-modal Metaphors to Browse Internet Search Results. In: *Department of Computing, School of Informatics. University of Bradford, Bradford* (2008)
17. Alotaibi, M.: *Electronic Customer Knowledge Management Systems: a Multimodal Interaction Approach*. In: *Informatics Research Institute. University of Bradford, Bradford* (2009)

Evaluation of WikiTalk – User Studies of Human-Robot Interaction

Dimitra Anastasiou¹, Kristiina Jokinen², and Graham Wilcock²

¹ University of Bremen, Germany

² University of Helsinki, Finland

dimitra@d-anastasiou.com,

{kristiina.jokinen, graham.wilcock}@helsinki.fi

Abstract. The paper concerns the evaluation of Nao WikiTalk, an application that enables a Nao robot to serve as a spoken open-domain knowledge access system. With Nao WikiTalk the robot can talk about any topic the user is interested in, using Wikipedia as its knowledge source. The robot suggests some topics to start with, and the user shifts to related topics by speaking their names after the robot mentions them. The user can also switch to a totally new topic by spelling the first few letters. As well as speaking, the robot uses gestures, nods and other multimodal signals to enable clear and rich interaction. The paper describes the setup of the user studies and reports on the evaluation of the application, based on various factors reported by the 12 users who participated. The study compared the users' expectations of the robot interaction with their actual experience of the interaction. We found that the users were impressed by the lively appearance and natural gesturing of the robot, although in many respects they had higher expectations regarding the robot's presentation capabilities. However, the results are positive enough to encourage research on these lines.

Keywords: Evaluation, multimodal human-robot interaction, gesturing, Wikipedia.

1 Introduction

In human-robot interaction (HRI) not only speech, but also other modalities, such as gesture and nodding make the conversation more natural, effective, and user-friendly. However, the evaluation of such intelligent agents requires a comprehensive setup to define correlations between different modalities and their usage, and to investigate human interaction and its basis as an *affordable* model for HRI. *Affordance* is a concept that was brought to HCI by [1] and refers to the properties that suggest to the user the appropriate ways to use the artifact. The concept's use for spoken dialogue systems was suggested by [2]: when interactive systems *afford* natural interaction techniques their interfaces lend themselves to a natural use without users needing to reason how the interaction should take place to get the task completed. [2] also points out the *rationality* of system actions meaning that the system is not regarded as a simple reactive machine or a tool, but capable of acting appropriately in situations which are not directly predictable from the previous action.

Concerning the evaluation of spoken dialogue systems, [3] stated that during the past years automatic evaluation and user simulations gained ground in order to enable quick assessment of design ideas without resource-consuming corpus collection and user studies. They distinguished between evaluation approaches (empirical vs. theoretical), conditions (laboratory vs. real usage), and goals, and categorized evaluation types in the following:

- Functional evaluation – Is the system functionally appropriate?
- Performance evaluation – How well does it perform the task it is designed for?
- Usability and quality evaluation – Are potential and real users satisfied with the performance, and how do the users perceive the system when using it?
- Reusability evaluation – Is the system flexible and portable?

In usability testing, questionnaires or subjective evaluations are used to learn from the users what a usable system is. To system designers, subjective evaluations may give more informative data of system functionality than objective performance measures, since they focus on the user's first-hand experience. For instance, recommendations for speech-based telephone services concern three different types of questionnaires:

- Those about the user's background at the beginning of an experiment.
- Those about the user's interactions with the system.
- Those about the user's overall impression at the end of an experiment.

The evaluation's goal and metrics are important factors to define how the different topics are translated into precise questions or statements. In our study we used two types of questionnaires: one at the beginning of the session to collect the user's expectations and one at the end of the experiment to collect the user's experience (cf. [10]).

The paper is laid out as follows: Section 2 presents the Nao WikiTalk application with focus on the gestures that were designed to enhance the robot's presentation and turn-management capabilities. Section 3 presents our user study and its setup, including instructions, the questionnaires and our methodology. Section 4 reports the results of our evaluation and Section 5 concludes the paper with some future prospects.

2 The Nao WikiTalk Application

WikiTalk [11] is a spoken dialogue system for open-domain knowledge access using Wikipedia as a knowledge source. Prototypes of WikiTalk were first developed using Windows Speech Engine and the Pyrobot robotics simulator [4]. The Nao WikiTalk version was implemented at the 8th International Summer Workshop on Multimodal Interfaces in Metz, France in July 2012. The Aldebaran Nao humanoid robot was used as the robot interlocutor in multimodal conversations. The work focused on different research issues, among others, on designing gestures, gaze tracking, integration of body motion with the spoken conversation system and also combination of suitable body language with Nao's own speech turns during the conversation. The main results of the joint work are reported in [5], while multimodal signaling is described in [6], and integration of gesturing and speech in [7].

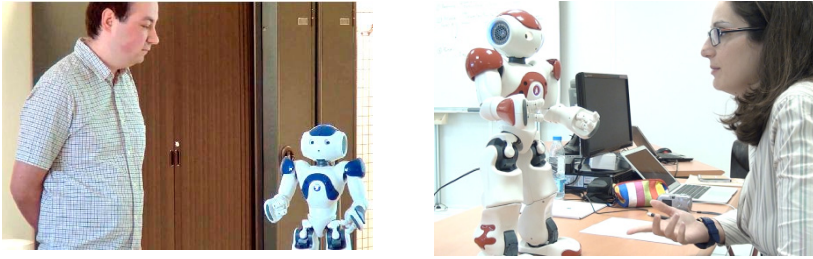


Fig. 1. Users interacting with the Nao robot

The Nao robot by Aldebaran Robotics (<http://www.aldebaran-robotics.com>) is a fully programmable humanoid robot, which has many sensors and actuators, and is of a convenient size and attractive appearance, with sophisticated embedded software (see Figure 1). Nao supports face and object recognition, speech recognition, text-to-speech, and whole body motion.

2.1 Gestures

Gesturing is a means of communication that can make HCI and HRI more natural, expressive, communicative, and user-friendly. A set of non-verbal gestures were designed in order to enhance Nao’s presentation and turn-management capabilities [7]. These apply Kendon’s [8] notion of gesture families. The *Open Hand Supine* (“palm up”) and *Open Hand Prone* (“palm down”) families have their own semantic themes related to giving ideas as well as presenting, explaining, summarizing vs. stopping and halting, respectively [9]. For the presentation capabilities, a set of presentation gestures were identified to mark the topic, the end of a sentence or paragraph, plus beat gestures and head nods to attract attention to hyperlinks (new information), and head nodding as backchannels (see more in [7]). For the turn-management capabilities, the following approach was applied: Nao speaks and observes the human partner; after each information chunk that Nao presents, the human is invited to signal continuation (phrases like ‘continue’ or ‘stop’); Nao asks explicit feedback depending on user’s turn; the robot may also gesture, stop, etc. depending on previous interaction. This shows the *rationality* of the system, i.e. situation-dependent appropriate actions of the robot.

3 User Study

In this section we discuss the study’s setup (3.1) as well as the questionnaires about the expectations and experience of the participants (3.2), and our methodology (3.3).

3.1 Setup

The user study took place at the 8th International Summer Workshop on Multimodal Interfaces (eNTERFACE 2012) in Metz. We ran user studies to test the application with 12 participants (5 female and 7 male). All participants were members of other projects organized by the summer school. They were in the age group 20-40 and came from various countries, including France, Germany, Switzerland, Greece, and India. The participants interacted with Nao in three phases/tests (5-10 minutes each phase).

The evaluation follows the framework of [10]. The users first answer a questionnaire concerning their expectations about the application, and then, after their experience of using it, they answer the same questions concerning their actual experience. The questions are the same, but their linguistic formulation is adapted to suit the future expectations and the past experiences accordingly.

Before starting the first test, an experimenter explained to the participant the tasks to be done in all the tests, gave a consent form to sign, and also handed an instruction sheet which the participants could take with them in another room where they have the interaction with Nao. Their task was to interact with Nao in an open conversation asking for a topic from Wikipedia and to try out how well it can present interesting information. Our instructions regarding the topics were the following:

- Nao can talk about almost any topic.
- You can change to another (related) topic simply by saying the name of one of the things that Nao mentions.
- You can interrupt Nao any time, by touching the front button on top of its head.
- You can move around and try to catch Nao's attention from different angles.
- You can finish the interaction session by saying *thank you*.

The robot suggests some topics to start with, and the user can shift to related topics by speaking their names after the robot mentions them. The user can also switch to a totally new topic by spelling the first few letters. The instruction sheet listed the main user commands to Nao ('continue', 'repeat', 'enough', etc.), as well as the spelling alphabet (A = Alpha [AL FAH], B = Bravo [BRAH VOH] etc.).

The evaluation by each user was divided in three sessions, lasting about 5-10 minutes each. Each session involved one of three different system versions and accordingly, the users had to evaluate three different interactions with Nao. Table 1 summarises differences between the different system versions. The first version did not include gesturing but only face tracking, while the two other versions differed in the number and variety of gestures and posture, to allow us to test the user reactions.

Table 1. Non-verbal gesture capabilities of Nao [8]

<i>System version</i>	<i>Exhibited non-verbal gestures</i>
System 1	Face tracking, always in the Speaking pose
System 2	Head Nod Up, Head Nod Down, Open Hand Palm Up, Open Hand Palm Vertical, Listening and Standing pose
System 3	Head Nod Up, Open Hand Palm Up and Beat Gesture (Open Hand Palm Vertical)

3.2 Questionnaires

Each participant filled in a questionnaire four times: first to give their expectations before starting their interaction with Nao, and then after each session to evaluate the system they had just interacted with. The questionnaire focused on the various aspects of the interaction and the robot's presentation. It included the following categories:

- a) *Interface*: sound, hand and body movements, face tracking.
- b) *Expressiveness*: instinctive, lively, natural way of communication.
- c) *Responsiveness*: speed of reaction, appropriate responses, easy to follow.
- d) *Usability*: easy to interrupt, easy to know what do next.
- e) *Overall*: head tracking/movements, enjoyment in interaction.

Each category included 5-7 statements, specific to the category. In this way, more detailed and accurate information from the user could be collected. The questionnaire was formulated as statements, and the user estimated how much they agreed with each statement on a 5-point Likert scale, from *I strongly agree* to *I strongly disagree*. An example of the *Interface* category can be seen in Table 2.

Table 2. Part of *expectations* questionnaire

I expect to understand quickly the purpose of the sounds emitted by Nao.
I expect to notice if Nao's hand gestures are linked to exploring topics.
I expect to find Nao's hand and body movement distracting.
I expect to find Nao's hand and body movements creating curiosity in me.
I expect to find Nao's face tracking speed appropriate.

Experience was measured with an analogous questionnaire after each user test, with the same categories. A sample of the *Interface* category is in Table 3.

Table 3. Part of *evaluations* questionnaire

I understood quickly the purpose of the sounds emitted by Nao.
I noticed Nao's hand gestures were linked to exploring topic.
Nao's hand and body movement distracted me.
Nao's hand and body movements created curiosity in me.
Nao's face tracking speed was appropriate.

There were also multiple choice questions, like *What do you expect will be the most difficult part of interaction?* At the end there were open-ended questions with free text answers for users to give comments on issues that had not been taken into account.

3.3 Methodology

There were in total 48 completed questionnaires: 12 about expectations and 36 (12 users and 3 evaluation tests) about experience. For the expectations, we calculated the

average of each category (*Interface*, *Expressiveness*, etc). Regarding the experience, we calculated the average for each category, but also for each evaluation test (Evaluation 1, 2, and 3). More importantly, we compared the expectations with the experience for each category and each evaluation test; for each of the 30 statements individually as well as the average scores. Furthermore, we ran paired t-tests to find out whether there is a statistically significant difference between the expectations and evaluations.

4 Results

4.1 User Expectations

Figure 2 shows the average expectations for the five evaluation categories.

The highest expectations fall under *Usability*, the lowest in *Expressiveness*, indicating that the users expected interaction to be functional, but not very natural. The single lowest expectation was in *Expressiveness* with a mean value of 2.75, and in

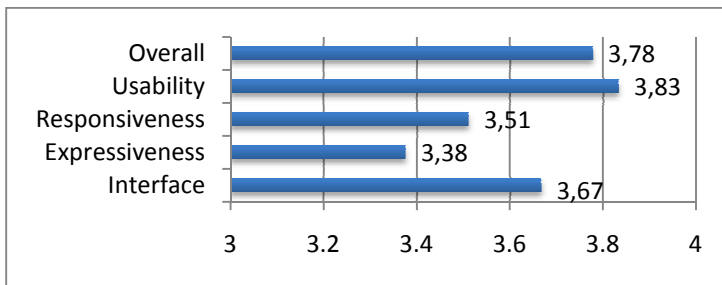


Fig. 2. Average of expectations in all categories (5-point Likert scale)

particular, the user did not expect Nao's gesturing to be natural. The highest expectation with a mean value of 4.25 was shared by the statements *I expect Nao's info to be correct* and *I expect to understand quickly the purpose of the sounds emitted by Nao*.

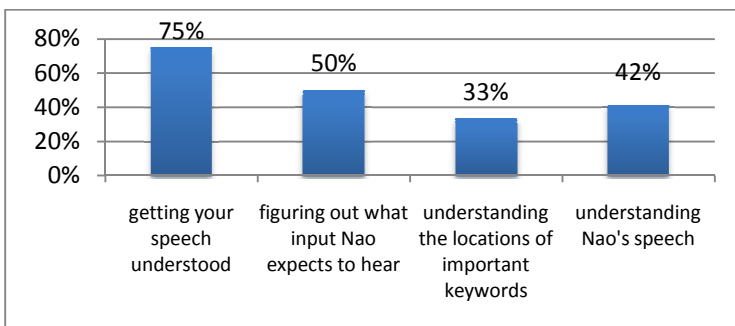


Fig. 3. Ranking of expectations about the most difficult part of interaction

Figure 3 shows the ranking of answers to the multiple choice questions concerning the most difficult parts of the interaction. Expectations about the difficulty of speech recognition (getting your speech understood) are significantly higher than those for speech synthesis (understanding Nao’s speech), 75% and 42%, respectively, reflecting perhaps the user’s prior knowledge of the problems in speech recognition. Participants were most confident (expected least problems) in that they would understand the locations of important keywords/topics that Nao can talk about (only 33% considered this a problem).

4.2 User Experience

Figure 4 compares the average of the expectations and of the evaluations.

As Figure 4 shows, in all categories (*Interface*, *Expressiveness*, etc.) the expectations were higher on average than the experience after the sessions. In categories *Interface*, *Expressiveness* and *Responsiveness*, System 2 was evaluated higher than the others (see Table 1 for the versions and gesture capabilities in each test). This gives support for the original goal by suggesting that the users appreciated and had a more positive experience of the interaction with full repertoire of gestures compared with the one with less expressive presentation. In *Usability*, System 3 was ranked slightly higher than the others, indicating perhaps that the users had become more familiar with how to interact with the system and thus experienced it more usable as well. In all categories, System 1 was ranked lowest, which supports our initial hypothesis that gesturing makes interaction with the robot more natural and expressive.

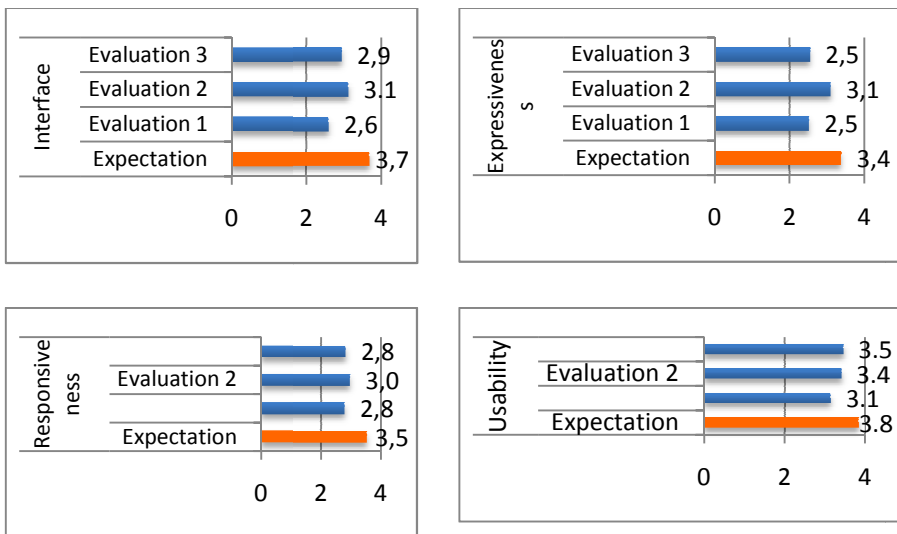


Fig. 4. Comparison between expectations and evaluations (5-point Likert scale)

Looking at individual statements, the highest scores were as follows. In *Interface*, the highest ranked statement in all three tests was *I understood quickly the purpose of the sounds emitted by Nao* (4.2, 4.3 and 4.2, respectively); this statement even exceeded the expectations in the second test slightly (4.2). In *Expressiveness*, the statements *Nao appeared lively* and *Nao's gesturing was natural* scored high in System 2 (3.8 and 3.4, respectively), and also exceeded user expectations (2.7 for both). These are positive results concerning our original goal of making the interaction more expressive by using gesturing. In *Responsiveness*, users ranked *Nao was slow to react* high in each test (3.7; 3.2 and 3.2, respectively), and also *It was easy to stop Nao speaking* (3.0, 3.6 and 3.4, respectively). It is interesting that the statement *Nao was able to change topic when I wanted* scored high in System 2 (3.4) and quite high also in System 3 (3.2), but not so in System 1 (2.8); it is likely that expressive gesturing also added to the user's positive experience of controlling the interaction. In *Usability*, the statement *I knew what to do when Nao stopped talking* was ranked the highest in System 3 (3.8).

4.3 Comparing Experience with Expectations

Figure 5 presents an overview of the evaluation categories. We can deduce from the curves that user's experience was similar to their expectations: *Usability* was ranked high and *Expressiveness* low. However, in *Expressiveness*, System 2 with full set of gestures was ranked much higher than the other two systems. In general, System 2 was experienced as the best of the three system versions, and scored closer to the expectations.

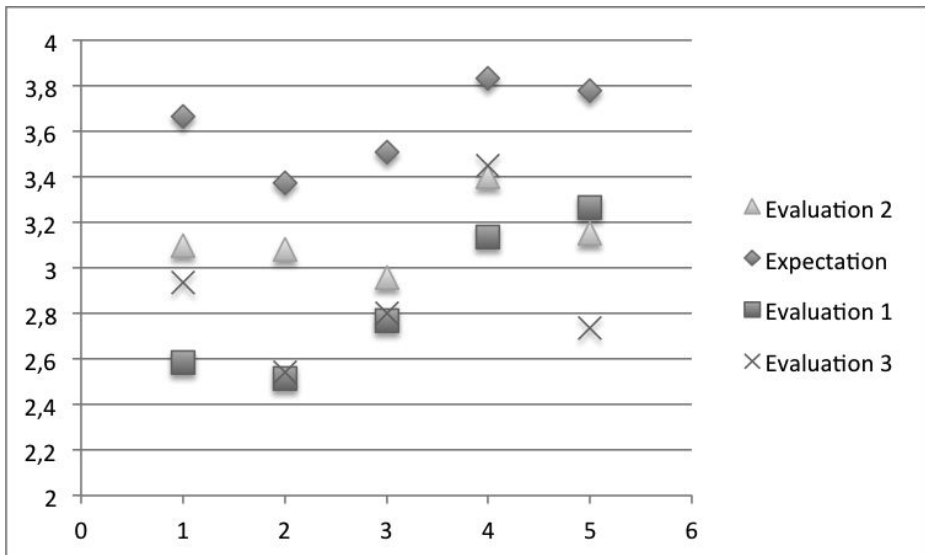


Fig. 5. Overall category and general overview

To test for a statistically significant difference between expectations and experience, we ran paired t-tests (alpha 0.05) on each evaluation category and the results are shown in Table 4. The boldfaced differences are significant on the level $p < 0.05$. We notice that with respect to System 1, the user’s expectations in all categories are significantly higher than their experience. With System 3, expectations are significantly higher than the experiences for *Expressiveness* and *Responsiveness*. However, with System 2, experience was significantly lower only with *Responsiveness*, while in other categories the user’s experience did not differ significantly from their expectations. It is interesting that the differences in *Expressiveness* of system 1 and *Responsiveness* concerning the Systems 2 and 3 are significant also on the stricter level of 0.01, marking these differences the most prominent ones among the system versions.

Table 4. Paired t-tests between user expectations along the different evaluation dimensions

Eval	Interface		Expressiveness		Responsiveness		Usability	
	t(4)	p	t(5)	p	t(7)	p	t(4)	p
1	3.34	0.029	5.09	0.004	2.9	0.022	3.4	0.027
2	1.21	0.290	0.77	0.475	4.5	0.003	2.3	0.079
3	1.80	0.146	3.26	0.022	5.1	0.001	1.6	0.174

We also compared the different evaluation versions with each other. In general, interactions with the different systems were not ranked very different from each other, but statistically significant differences could also be found. Interactions with System 2 were significantly more *expressive* than those with System 3, supporting the claim that a full repertoire of multimodal signals makes the interaction more natural and expressive. Evaluations 2 and 3 were significantly more *usable* than evaluation 1 and evaluation 2 was significant even on a tighter level of 0.01.

4.4 Free User Comments

In the evaluation questionnaires, apart from the statements, there was also a text box to fill in any comments that participants may have that were not covered in any of the statements. There were 23 comments in total and they fall into three categories: 9 comments were about speech or sound interaction; 3 were about interaction based on gestures and 3 were about the selection of topics; the remaining were without useful information (thanks, no comment, etc.). As far as speech interaction is concerned, it was commented that it was very laborious and should be faster and more accurate.

Many comments were related to the topic selection. Some noteworthy comments are that it was difficult to identify the topics that Nao mentions and participants wished that they had a list of topics. Another participant said that it looked like a constrained topic. In fact, participants could select another topic, but this was obviously unclear to them. Some of them commented on the sound interaction, i.e. it would be better not to have to wait for a beep before accepting human input (turn-taking), to reduce delay in the interaction, and that they had to speak close to make the robot understand. Regarding gestures, they mentioned that the arm movements came in arbitrary places during the conversation and should be clearer when they occur.

5 Conclusion and Future Prospects

In this paper we have described the evaluation of a robot application, Nao WikiTalk, and presented results comparing the user's expectations and experiences with respect to three different versions of the robot behaviour. The results show that System 2 with most human-like and affordable presentation of information was most highly valued and exceeded the user expectations in two respects: lively appearance and natural gesturing. The current prototype version supports multimodal interaction technology and provides a platform for experimenting with different interaction possibilities. In the future we plan to enhance Nao WikiTalk further with respect to its communicative capabilities, using more expressive and accurately timed gestures, and we will also focus on issues related to topic management and coherence of interaction.

Human-robot interaction is a fast growing and interesting research area, inviting deeper investigations into interaction between humans and intelligent agents. Exciting research topics include multiparty interaction where the robot is one of several interactive participants, and extending interactive situations into virtual environments. This research supports the development of services and applications that can improve daily life by providing more natural access to digital information.

Acknowledgments. We would like to thank the other project members Adam Csapo, Emer Gilmartin, Jonathan Grizou, Frank Han, and Raveesh Meena for their collaboration during the project. We would also like to thank the organisers of eNTERFACE-2012 for providing us with an excellent working environment.

References

1. Norman, D.A.: *The Psychology of Everyday Things*. Basic Books, New York (1988)
2. Jokinen, K.: Rational communication and affordable natural language interaction for ambient environments. In: Lee, G.G., Mariani, J., Minker, W., Nakamura, S. (eds.) *IWSDS 2010*. LNCS, vol. 6392, pp. 163–168. Springer, Heidelberg (2010)
3. Jokinen, K., McTear, M.: *Spoken Dialogue Systems*. Synthesis Lectures on Human Language Technologies, vol. 2(1). Morgan & Claypool (2009)
4. Jokinen, K., Wilcock, G.: Constructive interaction for talking about interesting topics. In: *Proceedings of Eighth Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, pp. 404–410 (2012)
5. Csapo, A., Gilmartin, E., Grizou, J., Han, J., Meena, R., Anastasiou, D., Jokinen, K., Wilcock, G.: Multimodal conversational interaction with a humanoid robot. In: *Proceedings of 3rd IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2012)*, Kosice, pp. 667–672 (2012)
6. Han, J., Campbell, N., Jokinen, K., Wilcock, G.: Integrating the use of non-verbal cues in human-robot interaction with a Nao robot. In: *Proceedings of 3rd IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2012)*, pp. 679–683. Kosice (2012)
7. Meena, R., Jokinen, K., Wilcock, G.: Integration of gestures and speech in human-robot interaction. In: *Proceedings of 3rd IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2012)*, pp. 673–678. Kosice (2012)

8. Kendon, A.: *Gesture. Visible Action as Utterance*. Cambridge University Press, Cambridge (2005)
9. Jokinen, K.: Pointing gestures and synchronous communication management. In: Esposito, A., Campbell, N., Vogel, C., Hussain, A., Nijholt, A. (eds.) *Second COST 2102. LNCS*, vol. 5967, pp. 33–49. Springer, Heidelberg (2010)
10. Jokinen, K., Hurtig, T.: User expectations and real experience on a multimodal interactive system. In: *Proceedings of 9th International Conference on Spoken Language Processing (Interspeech 2006)*, Pittsburgh (2006)
11. Wilcock, G.: WikiTalk: a spoken Wikipedia-based open-domain knowledge access system. In: *Proceedings of the COLING-2012 Workshop on Question Answering for Complex Domains*, Mumbai, pp. 57–69 (2012)

Robust Multi-Modal Speech Recognition in Two Languages Utilizing Video and Distance Information from the Kinect

Georgios Galatas^{1,2}, Gerasimos Potamianos^{3,2}, and Fillia Makedon¹

¹ Heracleia Human Centered Computing Lab,
Computer Science and Engineering Dept.,
University of Texas at Arlington, USA

² Institute of Informatics and Telecommunications,
NCSR "Demokritos", Athens, Greece

³ Dept. of Computer and Communication Engineering,
University of Thessaly, Volos, Greece

georgios.galatas@mavs.uta.edu, gpotam@ieee.org,
makedon@uta.edu

Abstract. We investigate the performance of our audio-visual speech recognition system in both English and Greek under the influence of audio noise. We present the architecture of our recently built system that utilizes information from three streams including 3-D distance measurements. The feature extraction approach used is based on the discrete cosine transform and linear discriminant analysis. Data fusion is employed using state-synchronous hidden Markov models. Our experiments were conducted on our recently collected database under a multi-speaker configuration and resulted in higher performance and robustness in comparison to an audio-only recognizer.

Keywords: Audio-visual automatic speech recognition, multi-sensory fusion, languages, linear discriminant analysis, depth information, Microsoft Kinect.

1 Introduction

Speech is the most natural form of communication for humans, and therefore automatic speech recognition (ASR) is one of the most intuitive forms of human-computer interaction (HCI). To improve ASR accuracy and robustness to noise, incorporation of visual information in conjunction with audio has been shown to be beneficial [1, 2]. However, in most research studies, such information is obtained from traditional planar video, thus not utilizing 3D visual speech articulation information. To alleviate this shortcoming, only a handful of efforts have appeared employing multiple or stereo cameras to capture the speaker's face [3-5], with an increase though in hardware cost and software complexity. We have recently proposed an alternative to such approach, by aiming to capture 3D visual speech information from the depth sensor of the novel Kinect device that operates based on the structured light method [6]. That work however considered audio-visual speech recognition (AVASR) in English only [7].

In this paper, we extend our previous work to consider AVASR in Greek, deviating from the traditional AVASR literature paradigm that considers one language only. Our system has been tested using two different languages, English and Greek, in a tri-stream multimodal fusion approach to ASR, where audio, planar video and distance information are combined for a small-vocabulary recognition task in order to keep data collection at a manageable level. Our experiments demonstrate consistent benefits when using the additional modalities to the performance and robustness for the ASR task across the two languages considered.

The design and experimentation using our system is presented in the next sections as follows: Initially, the system architecture is presented in Section 2 with details about the visual feature extraction and fusion. The experimental setup and results are discussed in Section 3 and our conclusions are presented in Section 4.

2 Description of the System Architecture

The input streams used by our system are audio, planar video and the distance information stream captured by the Kinect. The audio stream was captured using a Zoom H4 external voice recorder exhibiting good directionality and frequency response at 16-bit, 44.1kHz, PCM format. The planar video (24-bit color, VGA resolution) and the distance information (11-bit, VGA resolution) were both captured using the Kinect. The system architecture is shown in figure 1, and the various modules of the system are described in more detail in the following paragraphs.

2.1 Visual Front-End

The visual front-end is responsible for detecting and tracking the mouth region of interest (ROI) from each video frame. A nested setup using 2 Viola-Jones detectors [8] is used to detect the face and mouth of the speaker respectively. The nested implementation minimizes the number of false mouth detections by only searching for a mouth if the face of the speaker has already been detected. In addition, the coordinates of the mouth bounding box are smoothed by a median filter in order to minimize abrupt movements due to false detections. The detected mouth ROI coordinates from the video stream are also used for extracting the mouth region from the distance information stream. Finally, the size of both ROIs is normalized to 64x64 pixels.

2.2 Feature Extraction and Selection

The next step is to extract meaningful features from all 3 streams. For the audio stream, the well known Mel frequency cepstral coefficients (MFCCs) are extracted using the “Hidden Markov Model Toolkit” (HTK) [9], reaching a dimensionality of 39 (including first and second derivatives). For the video and distance streams, the coefficients of the 2-D discrete cosine transform (DCT) are extracted and interpolated to 100 Hz to match the rate of the audio features. Following is the feature selection step, which is comprised of 2 parts. Initially, the 45 highest energy coefficients of the upper left corner of each DCT image are selected as those with the highest information content. Subsequently, linear discriminant analysis (LDA) is applied to the features and those corresponding to the highest eigenvalues are selected as the most

informative ones. Finally, the first and second derivatives are appended to the final feature vector in order to capture the dynamics of speech. The final feature dimensionality is 21 for each of the video and distance streams.

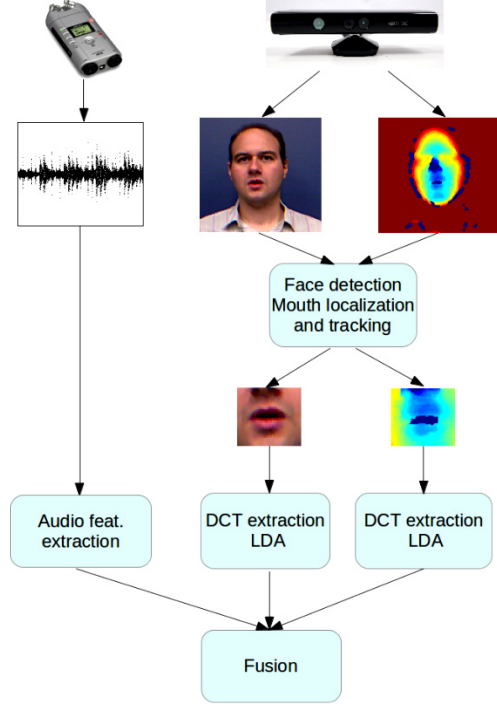


Fig. 1. Overview of the system modules

2.3 Data Fusion and Modeling

Hidden Markov models (HMMs) are the most commonly used classifiers for modeling speech. In our experiments we utilized state-synchronous multi-stream HMMs in order to effectively fuse the data from all 3 streams.

$$\Pr[o_t^{AVD} | c] = \prod_{s \in \{A,V,D\}} \left[\sum_{k=1}^{K_{sc}} \omega_{sck} N_{d_s}(o_t^{(s)}; m_{sck}, s_{sck}) \right]^{\lambda_{sct}} \quad (1)$$

This type of model realizes a decision-fusion approach, by computing the state emission (class conditional) probability as a product of the observation likelihoods of every stream, raised to a specific exponent λ , as shown in eq. 1. This exponent is bound to the reliability of the stream itself and defines the contribution of each stream. o_t^{AVD} denotes the tri-modal observation vector $o_t^{AVD} = \{o_t^A, o_t^V, o_t^D\}$, s is one of the three streams, c denotes the HMM state and t is the time (frame) of the utterance. The HMMs used in our work have a 3-state left-to-right topology, modeling

tri-phones with 16 Gaussian mixtures per stream and state. HTK patched with HTS [10] were used for training and testing using the aforementioned models.

3 Experimental Results and Discussion

To support this work, we have captured our own database, the bilingual audio-visual corpus with depth information or BAVCD [11], that includes audio, planar video, and distance measurements using a voice recorder, the Kinect, and an HD camera. The corpus contains data from 15 speakers for the English part and 6 speakers for the Greek part, uttering connected digit strings.

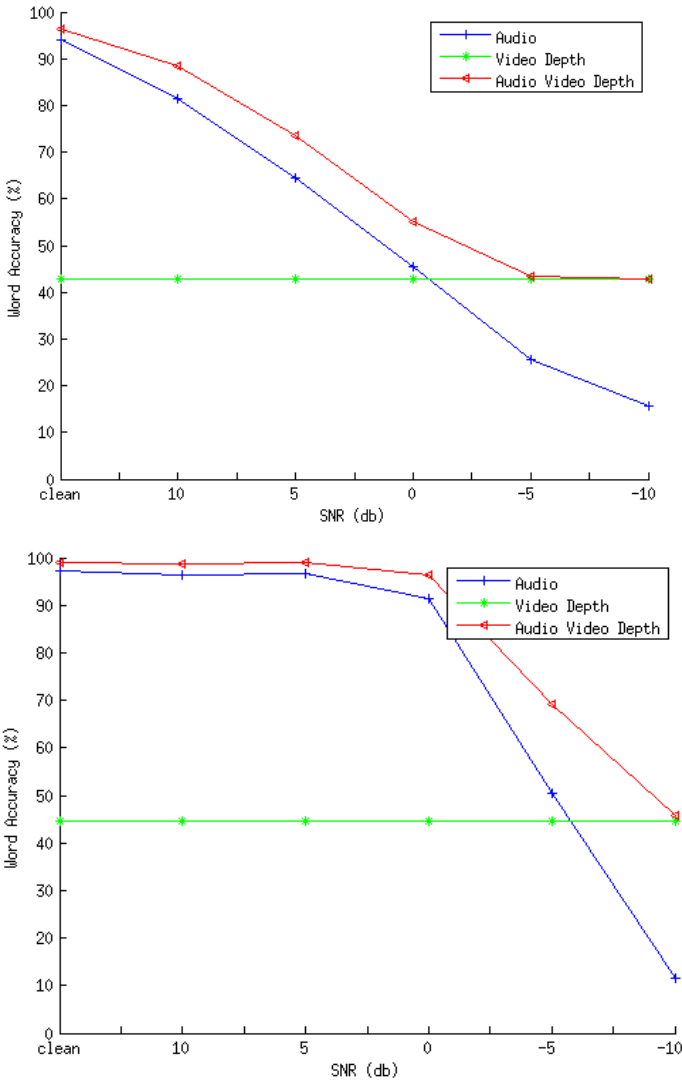


Fig. 2. Word accuracy results for English (top) and Greek (bottom) under various SNR levels

Our system was extensively tested in a multi-speaker setup under a variety of babble noise levels from the Noisex-92 database [12] in order to simulate a realistic smart home environment. Furthermore, training was conducted on clean speech, simulating the mismatch between the training and testing conditions. The results for each language are shown in figure 2. The best performance was achieved in the lack of noise when all 3 streams were utilized by the system. Under these conditions, the word accuracy for the Greek part was 99.02% and for the English part 96.32%. The performance for lip-reading without using audio was consistent for both parts and exhibited a 9.2% relative improvement using LDA. The overall system performance, degraded as the audio noise levels grew higher, but always remained higher than the performance of the individual streams, leading to a significant improvement under very noisy conditions e.g. 45.63% instead of 11.55% word accuracy for our system in comparison to an audio-only recognizer for a signal to noise ratio (SNR) of -10dB in Greek. The system exhibited better performance for the Greek language due to the smaller number of Greek speakers in the database in conjunction with the multi-speaker setup of the experiments.

4 Conclusions

In conclusion, we developed a novel speech recognition system that in addition to audio utilizes planar video and distance measurements captured by the Kinect and we tested its performance in both English and Greek. We have shown that our system exhibits high recognition rates in clean audio conditions but is also robust in noisy conditions, achieving significantly higher performance than an audio-only ASR system. Finally, our system's performance is consistent in both languages, constituting a reliable solution for speech recognition.

Acknowledgments. This material is based upon work supported by the National Science Foundation under Grants No. NSF-CNS 1035913, NSF-CNS 0923494. G. Potamianos would like to acknowledge partial support from the European Commission through FP7-PEOPLE-2009-RG-247948 grant AVISPIRE.

References

1. Iwano, K., Tamura, S., Furui, S.: Bimodal speech recognition using lip movement measured by optical-flow analysis. In: Proc. HSC, pp. 187–190 (2001)
2. Nakamura, S., Ito, H., Shikano, K.: Stream weight optimization of speech and lip image sequence for audio-visual speech recognition. In: Proc. ICSLP, vol. 3, pp. 20–24 (2000)
3. Goecke, R., Millar, B.: The audio-video Australian English speech data corpus AVOZES. In: Proc. ICSLP, vol. 3, pp. 2525–2528 (2004)
4. Vorwerk, A., Wang, X., Kolossa, D., Zeiler, S., Orglmeister, R.: WAPUSK20 – a database for robust audiovisual speech recognition. In: Proc. LREC (2010)
5. Ortega, A., Sukno, F., Lleida, E., Frangi, A., Miguel, A., Buera, L., Zacur, E.: AV@CAR: A Spanish multichannel multimodal corpus for in-vehicle automatic audio-visual speech recognition. In: Proc. LREC., vol. 3, pp. 763–767 (2004)

6. The Primesensor Reference Design, <http://www.primesensor.com>
7. Galatas, G., Potamianos, G., Makedon, F.: Audio-visual speech recognition incorporating facial depth information captured by the Kinect. In: Proc. EUSIPCO, pp. 2714–2717 (2012)
8. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2006)
9. Young, S.J., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P.: The HTK Book version 3.4. Cambridge University Press (2006)
10. The HMM-based speech synthesis system (HTS), <http://hts.sp.nitech.ac.jp>
11. Galatas, G., Potamianos, G., Kosmopoulos, D., Mcmurrough, C., Makedon, F.: Bilingual corpus for AVASR using multiple sensors and depth information. In: Proc. AVSP, pp. 103–106 (2011)
12. Varga, A., Steeneken, H.: Assessment for automatic speech recognition: Noisex-92. A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication* 12(3), 247–251 (1993)

The Ecological AUI (Auditory User Interface) Design and Evaluation of User Acceptance for Various Tasks on Smartphones

Myounghoon Jeon¹ and Ju-Hwan Lee²

¹ Michigan Technological University, Houghton, Michigan, USA
mjeon@mtu.edu

² Korean German Institute of Technology, Seoul, South Korea
jhlee@kgit.ac.kr

Abstract. With the rapid development of the touch screen technology, some usability issues of smartphones have been reported [1]. To tackle those user experience issues, there has been research on the use of non-speech sounds on the mobile devices [e.g., 2, 3-7]. However, most of them have focused on a single specific task of the device. Given the varying functions of the smartphone, the present study designed plausibly integrated auditory cues for diverse functions and evaluated user acceptance levels from the ecological interface design perspective. Results showed that sophisticated auditory design could change users' preference and acceptance of the interface and the extent depended on usage contexts. Overall, participants gave significantly higher scores on the functional satisfaction and the fun scales in the sonically-enhanced smartphones than in the no-sound condition. The balanced sound design may free users from auditory pollution and allow them to use their devices more pleasantly.

Keywords: Auditory user interface, ecological user interface design, smartphones, user acceptance.

1 Introduction

Smartphones have become one of the most important necessities in our daily lives. However, with the rapid development of the touch screen technology, some usability issues of smartphones have been reported [1]. Because smartphones and other touch screen devices usually have overlapped control and visual display areas and lack tactile feedback, the appropriate use of sounds could not only provide solutions to usability issues, but also enhance user experience [7]. Research [8] has supported this notion by showing that auditory feedback is the most effective modality in physical user interface satisfaction, followed by tactile and motion feedback.

Over the last two decades, there has been much research on the use of non-speech sounds to improve user interaction on the mobile devices [e.g., 2-7], but most of them have focused on a specific function or task of the device. It is not easy to find out literature that provides guidelines on how to lay out or integrate various auditory cues for complex functions depending on usage contexts. The present study attempts to

design and evaluate a more integrated auditory user interface set to better represent the use of smartphones with sounds in varied situations from the perspective of the ecological interface design [9].

1.1 Types of Non-speech Auditory Cues on Mobile Devices

There are a number of types of non-speech auditory cues that have been applied to touch screen devices. Auditory icons [10] use representative part of sounds of objects, functions, and events which bear an analogic relationship with the object they represent. For example, Wilson and his colleagues employed auditory icons in their SWAN system [11], a system for wearable audio navigation for blind people. Earcons (“ear + icons”) [12], on the other hand, use a short musical motive as symbolic representations of actions or objects. Earcons have shown superior performance compared to other less systematic sounds or no sound in a PDA [5] or a mobile phone [3]. Some researchers have also provided empirical guidelines for better aesthetics in creation of earcons [3, 13]. Auditory scrollbars also use musical sounds to represent a location of a user in a display (i.e., contextual information) as an analogy of visual scrollbars, which could be found on a computer application or a smartphone [14]. That previous study focused on a continuous scrollbar (the thumb could be located anywhere along the bar, whereas Yalla and Walker [15] examined the possibility of the use of discrete auditory scrollbars (the thumb can only be located at discrete, designated points) in mobile device menus. Yalla and Walker demonstrated the potential benefits of the proportionally mapped auditory scrollbars for visually impaired participants as well as sighted people. Brewster [14] once implemented a sonically enhanced widget set including buttons, menus, scrollbars, alert boxes, windows, and drag and drop on the desktop computer, but it focused on the use of only earcons and auditory scrollbars. Fairly recently, musicons (“music + icons”) have been introduced to the HCI community [16]. Musicons are brief samples of well-known music used in the auditory interface design. Researchers provided some usage scenarios such as the use of musicons as a reminder at home and preliminary guidelines to create better musicons. However, a further validity test is needed for the relationship between musicons and their intended meanings in the interface, in addition to the discernability of musicons as original songs. As relatively new auditory cues that combines speech and non-speech sounds, spearcons [“speech earcons” 17] and spindex [6] were introduced into mobile devices to overcome the shortcomings of either purely non-speech sounds (auditory icons) or music-driven auditory cues (earcons, auditory scrollbars, and musicons). Spearcons use compressed speech which is produced by speeding up spoken phrases [17]. These unique sounds blend the benefits of speech and non-speech because of the acoustic relationship between the spearcons and the original speech phrases. The use of spearcons has enhanced navigational efficiency as well as subjective satisfaction on the various mobile devices [e.g., 4, 18]. A spindex [“speech index” 6] uses a short cue created based on the pronunciation of the first letter or phoneme of each spoken menu item. For instance, the spindex cue for “Super” would sound /es/ or even /s/. The set of spindex cues in an alphabetical auditory menu is analogous to the visual index tabs in a reference book (e.g., a large dictionary). The use

of spindex cues has shown promising results in a mobile phone addressbook [6], on a touch screen device with various gesture styles [7], or in a dual task context [19].

To summarize, auditory display researchers have tried to use diverse non-speech auditory cues in mobile devices and have yielded successful outcomes as shown above. However, given the growing and varying functions and structures on the smartphone interface, more integrated research on the application of the sounds is needed for implementing sophisticated, but balanced auditory user interfaces.

1.2 Ecological User Interface Design

An ecological interface design (EID) approach focuses more on environments rather than on a specific task in complex and complicated systems [9]. This interface design framework aims to lessen mental workload and thus, enable users to more easily acquire advanced mental models about the system, by focusing on its entire architecture [20]. To this end, in the EID framework, various methods [e.g., abstraction hierarchy 21] have been used to determine what types of information should be displayed on the system interface and how the information should be arranged on different abstract levels. Given that environmental factors (e.g., noise, mask, interference) and an overall usage flow (e.g., harmony with one another) are as important in auditory displays as (or even more important than) in visual displays, the ecological interface design approach to auditory displays has been proposed and supported [e.g., 22, 23]. However, there has been rare empirical research on this topic.

From these backgrounds, the present study investigated users' acceptance of the plausibly integrated various auditory cues for specified purposes on a single touch screen smartphone. After identifying functions which auditory cues could be applied to through a preliminary function analysis of the target device, professional sound designers created several alternative sounds for each function. For all the task scenarios, we measured user acceptance levels and finally measured overall user experience of the presence of the sound. Based on this attempt, we expect that we could come up with a blueprint of the optimal layout of the auditory cues on the smartphone.

2 Method

2.1 Participants

Forty six (under) graduate students participated in this study (mean age = 23.4; female 27, male 19). All reported normal or corrected-to-normal vision and hearing, signed informed consent forms, and provided demographic details about age and gender.

2.2 Apparatus and Stimuli

Stimuli were presented using an LG LH 2300, smartphone with a 3 inch resistive wide full touch screen panel. The internal sound chip was used for sound rendering.

Participants listened to auditory stimuli using Sennheiser HD 202 headphones plugged into the phone's audio jack, and adjusted for fit and comfort.

Based on literature reviews on the guidelines of the sound applications to the electronic devices [24-26] and our function analysis of the target smartphone, we categorized generic functional groups to which sounds can be applied to facilitate user interaction as follows: (1) touch feedback, (2) sound widgets (3) hierarchical menu navigation, and (4) list menu navigation. A professional user interface designer, a sound designer, and an academic researcher iteratively discussed together and created all of the sound designs for this experiment. (1) Touch feedback: We focused on two types of touch tasks. The one was a flip, which is a pervasive gesture on the smartphone nowadays. The flip tones included a single tone, which provided a single feedback sound about users' flip action and revolving tones, which provided feedback sounds whenever each item passes by a specific zone until the rolling stops. These designs were made up of mechanical sounds (just like the wheeling sound on the Apple iPod), instead of an instrumental sound. The other touch feedback task was the dialing tone on the phone. The dialing tones included a single tone (the original sound of that smartphone), double tones (applying different sounds to "touch" and "release" actions each), and a voice (speaking out the touched number in addition to a single tone). (2) Sound widgets: For sound widgets, six different functions were included: an auditory scrollbar [e.g., 14, 15], auditory progress bar [e.g., 27] and auditory progress circulation, auditory check box, auditory toggle, and auditory pop-up. Auditory scrollbars and progress bars were studied before, but we extended the previous design and added alternatives. For the remaining functions, we devised new alternative ones. Two different auditory scrollbars were created. The one was the effect sound, which lasted by proportional length based on the distance that the scrollbar moved. Another was the location mapping sound, which was composed of four different notes. The first two notes presented the possible entire range of the scrollbar (e.g., if it goes down, top and bottom notes). The last two notes presented the current location of the scrollbar out of the range. (e.g., if it goes down and is in the middle of it, middle and bottom notes) The polarity of the sounds was changed depending on the moving direction of the scrollbar. The auditory progress bars involved three different designs. The first was the step sound, which got gradually faster as approaching the end. The second was the spatial sound, which was composed of one pitch, moving from left to right. The third was the pitch change, in which incremental tones with higher pitch were added as time goes by. For the auditory progress circulation, we created a single sound design, which was composed of a psychologically circulating sound called a 'Shepard tone'. The auditory check box sounds consisted of two alternatives. One was the single tone, which was heard whenever a user checks on a checkbox. Another was the selected number tone, in which the user could hear a total number of currently checked boxes whenever the user adds or subtracts a check mark (e.g., if a user checks the second checkbox, then he or she will hear two tones. If the user minuses a checkbox from the three marked checkboxes, he or she will also hear two tones because the currently marked boxes will be two). Two auditory toggle sounds were composed. In the first design, a single sound was generated whenever a user touches the toggle button. In the second design,

if the toggle button had three different modes, it generated three different (but structurally similar) sounds (e.g., C-E-G, C-G-E, G-C-E). Four different pop-up sounds were devised. The first one was just a single tone that notified that ‘there is a pop-up window’. The second was the continuous effect sound, which generated the sound intermittently until the user touches an OK button. The third was the simple repeated sound, in which a single tone was continuously repeated. The last one was the continuous melody, in which a melodious instrumental sound held out. (3) For the hierarchical menu navigation, sounds were tested in the two different functions, the menu depth and the menu content. The ambient menu depth cues were composed of the hierarchical background sound and unique background sound. The former included the same sound in the same depth, regardless of the functional category (e.g., if the menu structure has three different levels, only three distinct sounds are used in each depth). In contrast, the latter included a unique sound for each functional group, reflecting each depth. In both designs, the deeper level included incrementally more complex sounds (e.g., suppose the 1st depth includes a hi-hat sound. Then, the 2nd depth includes hi-hat and bass drum sounds. The 3rd depth includes hi-hat, bass drum, and snare drum sounds, which would provide navigation crumbs). The menu content auditory icons in each depth contained three different versions of sounds. The first one was a simple default sound. The second was the serial auditory icons that generated a couple of representative sounds of the functions in that category in a serial manner (e.g., generating a game sound and a camera shutter sound one by one when selecting the multimedia menu). The last one was the parallel auditory icons that generated a couple of representative sounds of the functions in that category in a parallel manner (e.g., generating a game sound and a camera shutter sound simultaneously). The serial auditory icons could clearly present functions, but they are relatively longer. In contrast, the parallel auditory icons could provide a quick auditory scan, but might be confusing. (4) List menu navigation: for the list menu navigation, two types of sounds were used. The first one was the single tone, which was the same as the single tone of the flip tone. Another sound condition used basic spindex cues. For instance, the spindex cue for “Michigan” would be a sound based on the spoken sound “M”.

2.3 Design and Procedure

A within-subjects design was used in this experiment. Thus, one participant experienced all the sounds and compared them. Our experiment was designed in this way to focus more on the intra-participant’s perception about overall integrated effects of various auditory cues on a single device, considering harmonization in between different tasks. After the consent form procedure, participants were seated in front of the desk on which they can play around with the smartphone. For each task, participants were instructed about the usage situation and how to control the smartphone and then, they conducted the task. The order of appearance of each task was counterbalanced across all participants. For every task, each condition was randomly presented to the participants. After completing each task, participants filled out the subjective

questionnaire on the ‘functional satisfaction’, the ‘fun’, and the ‘preference’ using a seven-point Likert-type scale. Finally, participants provided comments on the study.

3 Results and Discussion

For all of the data analysis, we used repeated measures analysis of variance (ANOVA) first. Then, for the pairwise comparisons, paired-samples t-tests were used with a conservative alpha level (.01) instead of .05, across all the comparisons.

3.1 Touch Feedback

- Flipping tone: Participants showed higher functional satisfaction ($M = 4.93$) and fun ($M = 4.74$) scores in the revolving sound condition than in the no sound ($M = 4.02$; $M = 2.67$) or the single tone ($M = 4.02$; $M = 3.26$) conditions ($ps < .01$). Participants tended to prefer the revolving sound condition ($M = 4.35$) to the single tone condition ($M = 3.61$), but this difference did not lead to a traditionally significant level ($p = .08$). Based on the results, auditory user interface designers could consider employing revolving-type sounds to the flip function on smartphones.

- Dialing tone: For the functional satisfaction scale, the single tone ($M = 4.78$) and the voice ($M = 5.04$) conditions showed significantly higher scores ($ps < .01$) than the no-sound ($M = 3.89$) or the double tone ($M = 4.22$) conditions. Nevertheless, the voice ($M = 2.80$) condition was significantly less preferred than the single tone ($M = 4.43$) condition ($p < .01$), which means that even though the use of voice might functionally help users, they would choose a simple, brief single tone feedback for dialing, given that the dialing is a frequently used function. According to the results, the voice tone could be included in the setting option.

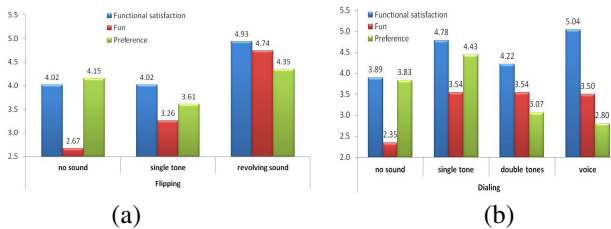


Fig. 1. Rating results about touch feedback (flipping (a) and dialing (b))

3.2 Sound Widgets

- Auditory scrollbar: Participants showed significantly higher scores on the functional satisfaction, the fun, and the preference scales in the location mapping sound condition than in the no sound or the effect sound conditions ($ps < .01$). Note that in the preference rating scale, the effect sound condition was significantly lower than the no sound condition ($p < .01$). The use of inappropriate sounds may make users

annoying.

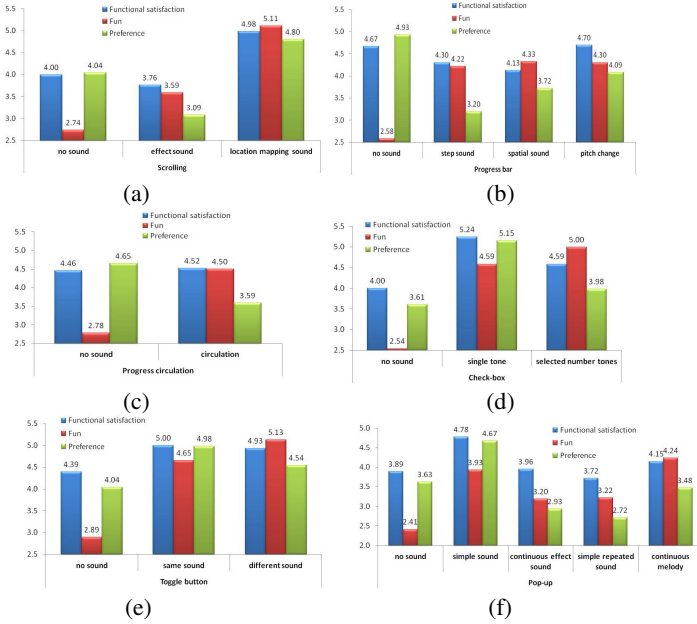


Fig. 2. Rating results about sound widgets (scrollbar (a) and progress bar (b), progress circulation (c), checkbox (d), toggle (e), and pop-up (f))

- Auditory progress bar and circulation: All the sound conditions showed significantly higher fun scores than the no sound condition ($ps < .01$), but showed lower preference scores. The pitch change condition showed significantly higher preference scores than the step sound or the spatial sound conditions ($ps < .01$), but it was still significantly lower than the no sound ($M = 4.93$) condition ($p < .01$). The results of the auditory progress circulations confirm this pattern. Whereas participants had fun in the circulation sound condition ($p < .01$) compared to the no sound condition, they preferred the no sound condition ($p < .01$) over the circulation sound. Previous research on auditory progress bars has shown that subjective workload can be lessened depending on the characteristics of the sounds [27]. However, in their experiment, there was no no-sound condition. Therefore, we may not over-generalize the results of auditory progress bars and circulations, but further research with more alternatives is needed.

- Auditory checkbox: Participants showed significantly higher functional satisfaction, fun, and preference in the single tone ($M = 5.24$; 4.59 ; 5.15) condition than in the no sound ($M = 4.0$; 2.54 ; 3.61) condition ($ps < .01$). For the fun scale, the selected number tone ($M = 5.0$) condition was also significantly higher than the no sound ($M = 2.54$) condition ($p < .01$). Moreover, given that all of the scores of the selected number tone condition were numerically higher than the no sound condition, sound designers could consider having this option in the setting menu even though they might want to have a single tone as a default.

- Auditory toggle: As to the auditory checkbox, participants favored the sound conditions over the no sound condition ($ps < .01$). The different sound condition showed the highest score ($M = 5.13$) on the fun scale, but the same sound condition showed the highest score on the functional satisfaction ($M = 5.0$) and the preference ($M = 4.98$) scales. There was no significant difference between the two sound conditions.
- Auditory pop-up: Participants preferred the simple notification pop-up sound to other conditions including the no sound condition ($ps < .01$). For the fun scale, the continuous melody condition was not significantly higher than the simple sound condition, but it showed numerically the highest score, which is promising for the next implementation. Improvements in terms of length, amplitude, or variations are expected to enhance functional satisfaction and preference of the continuous melody pop-up sound.

3.3 Hierarchical Menu Navigation

- Ambient menu depth cue: Participants showed significantly higher scores on the functional satisfaction and the fun scales in the unique menu depth background sound ($M = 4.46; 4.70$) condition than the hierarchically same depth background sound ($M = 4.0; 3.76$) condition ($ps < .01$). The unique sound score ($M = 3.80$) was also numerically higher than the hierarchically same sound ($M = 3.50$). Overall, using a specific sound theme for each menu category could make more functionally helpful and fun user interfaces on smartphones.
- Menu content auditory icon: Participants showed significantly higher scores on the functional satisfaction and the fun scales in three different sound conditions than the no sound condition ($ps < .01$). For all of the measures, participants gave numerically higher scores to the serial auditory icon than the parallel auditory icon, which means that complexity matters. Even though the serial auditory icon should be intrinsically longer than the parallel auditory icon, participants favored hearing one at a time over hearing different sounds simultaneously.

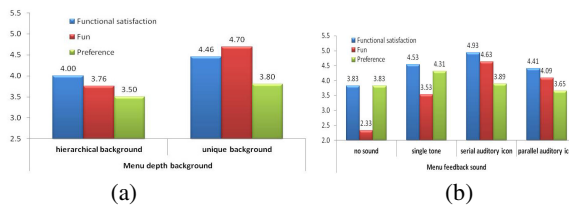


Fig. 3. Rating results about hierarchical menu navigation (ambient menu depth cue (a) and Menu content auditory icon (b))

3.4 List Menu Navigation

Participants showed significantly higher scores on the functional satisfaction and the fun scale in the spindex list-search condition than the no sound or the single tone conditions ($ps < .01$). However, the preference was not different. The subjective

preference could be improved using alternative designs of the spindex (e.g., the attenuated version or decreased version) [6].

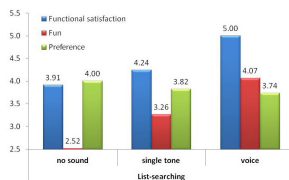


Fig. 4. Rating results about list menu navigation

4 Overall User Experience and Conclusion

For the overall user experience of the device, participants showed significantly higher functional satisfaction and fun scores in our sonically-enhanced smartphones than the no-sound condition ($ps < .01$). However, there was no preference difference. Based on overall results, sound designers could decide which sounds should be used as a default or as an option. Seven participants mentioned that the use of sound is going to be useful for the blind, the elderly, or children. More than half participants (26) stated that a variety of sounds are useful and functional, but simultaneously they mentioned that sounds should be turned off as an option. Thirteen participants said, “sounds are fun”, but eight participants used a silent mode as a default on their phone. An appropriate application of sounds does not guarantee that users, who used to turn off their sound mode, would turn on. However, the balanced sound design might free users from auditory pollution and allow users to use smartphones more pleasantly.

References

1. Norman, D.A., Nielsen, J.: Gestural interfaces: A step backward in usability. *Interactions* 17(5), 46–49 (2010)
2. Brewster, S.A., Leplâtre, G., Crease, M.G.: Using non-speech sounds in mobile computing devices. In: *Proceedings of the 1st Mobile HCI 1998*, Glasgow, UK (1998)
3. Leplâtre, G., Brewster, S.A.: Designing non-speech sounds to support navigation in mobile phone menus. In: *Proceedings of the 6th ICAD 2000*, Atlanta, GA, USA (2000)
4. Palladino, D., Walker, B.N.: Efficiency of spearcon-enhanced navigation of one dimensional electronic menus. In: *Proceedings of the ICAD 2008*, Paris, France (2008)
5. Brewster, S.A., Cryer, P.G.: Maximising screen-space on mobile computing devices. In: *Proceedings of the ACM CHI 1999*, Pittsburgh, PA, USA (1999)
6. Jeon, M., Walker, N.B.: “Spindex” (Speech Index) improves acceptance and performance in auditory menu navigation for visually impaired and sighted users. *ACM Transactions on Accessible Computing* 3(3), 10:1–26:1 (2011)
7. Jeon, M., Walker, B.N., Srivastava, A.: “Spindex” (speech index) enhances menu navigation on touch screen devices with tapping, wheeling, and flicking gestures. *ACM Transactions on Computer-Human Interaction* 19(2), 14:1–27:1 (2012)

8. Oh, J.W., Park, J.H., Jo, J.H., Lee, C., Yun, M.H.: Development of a kansei analysis system on the physical user interface. In: Proceedings of the HCI, Kangwon, Korea (2007)
9. Rasmussen, J., Vicente, K.J.: Coping with human errors through system design: Implications for ecological interface design. *International Journal of Man-Machine Studies* 31, 517–534 (1989)
10. Gaver, W.W.: Auditory icons: Using sound in computer interfaces. *Human-Computer Interaction* 2, 167–177 (1986)
11. Wilson, J., Walker, B.N., Lindsay, J., Cambias, C., Dellaert, F.: SWAN: System for wearable audio navigation. In: Proceedings of the 11th ISWC 2007 (2007)
12. Blattner, M.M., Sumikawa, D.A., Greenberg, R.M.: Earcons and icons: Their structure and common design principles. *Human-Computer Interaction* 4, 11–44 (1989)
13. Helle, S., Leplatre, G., Marila, J., Laine, P.: Menu sonification in a mobile phone – a prototype study. In: Proceedings of the ICAD 2001, Espoo, Finland (2001)
14. Brewster, S.A.: The design of sonically-enhanced widgets. *Interacting with Computers* 11(2), 211–235 (1998)
15. Yalla, P., Walker, B.N.: Advanced auditory menus: Design and evaluation of auditory scroll bars. In: Proceedings of the ASSETS 2008 (2008)
16. McGee-Lennon, M., Wolters, M.K., McLachlan, R., Brewster, S., Hall, C.: Name that tune: Musicons as reminders in the home. In: Proceedings of the CHI 2011, BC, Canada (2011)
17. Walker, B. N., Lindsay, J., Nance, A., Nakano, Y., Palladino, D. K., Dingler, T., Jeon, M.: Spearcons (speech-based earcons) improve navigation performance in advanced auditory menus. *Human Factors* (2012) (Published online July 2, 2012 Print edition pending)
18. Walker, B.N., Kogan, A.: Spearcon performance and preference for auditory menus on a mobile phone. In: Stephanidis, C. (ed.) UAHCI 2009, Part II. LNCS, vol. 5615, pp. 445–454. Springer, Heidelberg (2009)
19. Jeon, M., Davison, B.K., Nees, M.A., Wilson, J., Walker, B.N.: Enhanced auditory menu cues improve dual task performance and are preferred with in-vehicle technologies. In: Proceedings of the AutomotiveUI 2009, Essen, Germany (2009)
20. Vicente, K.J.: Ecological interface design: Supporting operator adaptation, continuous learning, distributed collaborative work. In: Proceedings of the HCPC (1999)
21. Rasmussen, J.: The role of hierarchical knowledge representation in decision making and system management. *IEEE Transactions on Systems, Man and Cybernetics* 15, 234–243 (1985)
22. Sanderson, P.M., Anderson, J., Watson, M.: Extending ecological interface design to auditory displays. CSIRO (2000)
23. Walker, B.N., Kramer, G.: Ecological psychoacoustics and auditory displays: Hearing, grouping, and meaning making. In: Neuhoff, J. (ed.) *Ecological psychoacoustics*, pp. 150–175. Academic Press, New York (2004)
24. Jeon, M.: Two or three things you need to know about AUI design or designers. In: Proceedings of the ICAD (2010)
25. Kurakata, K., Mizunami, T., Yomogida, H.: Guidelines on the temporal patterns of auditory signals for electronics home appliances: Report of the association for electric home appliances. *Acoust. Sci. & Tech.* 29(2), 176–184 (2008)
26. Lee, J.-H., Jeon, M., Han, K.H.: Developing the design guideline of auditory user interface for domestic appliances. *Korean Journal of the Science of Emotion & Sensibility* 10(3), 307–320 (2007)
27. Kortum, P., Peres, S.C., Stallmann, K.: Mental workload measures of auditory stimuli heard during periods of waiting. In: Proceedings of the HFES (2010)

Speech-Based Text Correction Patterns in Noisy Environment

Ladislav Kunc, Tomáš Macek, Martin Labský, and Jan Kleindienst

IBM Prague R&D Lab
V Parku 2294/4, 148 00 Prague 4
{ladislav_kunc1,tomas_macek,martin.labsky,
jan_kleindienst}@cz.ibm.com

Abstract. We present a study focused on observation of methods of dictation and error correction between humans in a noisy environment. The purpose of this study is to gain insight to natural communication patterns which can then be applied to human – machine interaction. We asked 10 subjects to conduct the standard Lane Change Test (LCT) while dictating messages to a human counterpart who had to note down the message texts. Both parties were located in separate rooms and communicated over Skype. Both were exposed to varying types and levels of noise, which made their communication difficult and forced the subjects to deal with misunderstandings. Dictation of both short and longer messages was tested. We observed how the subjects behaved and we analyzed their communication patterns. We identified and described more than 20 elementary observations related to communication techniques such as synchronization and grounding of parties, error checking and error correction. We also report frequencies of use for each communication pattern and provide basic characteristics of driving distraction during the test.

1 Introduction

Various communication, navigation and entertainment systems are becoming a part of our everyday in-car experience. Managing them without compromising safety becomes an issue. Appropriate choice of user interface techniques tailored for in-car use can be a critical factor in minimizing driving distraction.

Although many in-car devices nowadays primarily use various forms of not yet standardized controls (rotary knobs, joysticks or touch screens), speech UIs seem to be very promising. Correct design of speech interfaces for cars is however a relatively complex task. It requires proper judgment of various aspects including limited short term memory of the user under cognitive load, finding the right way how to benefit from other modalities, allowing fast recovery after UI interaction had to be interrupted by handling road situation and many other aspects. As part of designing a dictation and error correction UI for a car (see e.g. ECOR [2]), one thing which comes to mind is whether we could benefit from systematically observing the ways how humans do the same task themselves. We summarize the knowledge gained during this study in this paper. Although the study does not bring any quite unexpected observations we

found collected experience very beneficial to our main, man-machine interface design task. It helped us gain a good insight into how human dialog is organized and what are the basic tasks and methods used.

Obviously, we do not imply here that communication patterns between the two humans are the same as in human to machine conversation. If the user is aware that the machine is “on the other end” s/he adapts the conversation to the limits of the machine. The purpose of this study is to learn natural communication patterns of the humans for the task of dictation in a car. Deeper understanding of it should help us in future UI design.

2 Related Work

Significant attention was devoted in the past to judging the impact of in-car activities [1]. The Lane change Test (LCT) [3] and subjective tests using questionnaires such as NASA TLX and DALI [6] are examples of popular methods used to assess the impact of various in-car tasks that are secondary to the primary task of driving.

Although electronics systems are more and more abundant in cars, which rightfully causes worries about their impact on driving, communication between the driver and passengers is frequent and hardly can be regulated [5].

Several approaches to designing speech-based UIs for in-car usage were described including menu-based and search-based UIs [8].

Various patterns of text entry correction were elaborated in desktop speech-based solutions [4]. This paper discusses similar strategies for the car environment.

3 Experimental Setup

We decided to collect communication patterns for a constrained scenario where the driver dictates pieces of text over a hands-free phone. The driver’s counterpart, the note-taker, simulates a person sitting on the passenger’s seat who communicates with the driver. Non-verbal communication was not considered and is out of the scope of this paper.

Tests were conducted in laboratory environment. The drivers were using a low-fidelity driving simulator to mimic driving on a highway. The primary task performed was the standard LCT. Driving statistics were recorded and they are reported in Section 6.1. Each driver communicated over Skype with a note-taker located in a different room. The participants did not have any visual contact. The note-taker typed the received text to a computer text editor.

3.1 Distraction by Noise

To introduce communication errors, we exposed both the driver and the note-taker to a pre-recorded noise audio which was mixed with Skype audio in real-time and the mixed result was playing in headphones for both parties.

One can expect (confirmed by tests) that behavior of the user depends on noise level and also on its character. During pilot experiments we used noise of a starting car followed by more quiet periods of running engine. This recording was played in a loop. We observed fast adaptation of the subjects to the character of the noise as they quickly learned to wait for the quiet periods. Although this is a valid usage pattern, real-life types of noise are frequently not of this kind and cannot be always predicted. Therefore the noise used in the experiments presented here was modified to contain a high variety of sounds. A fragment of the utilized noise track is depicted by Fig. 1.

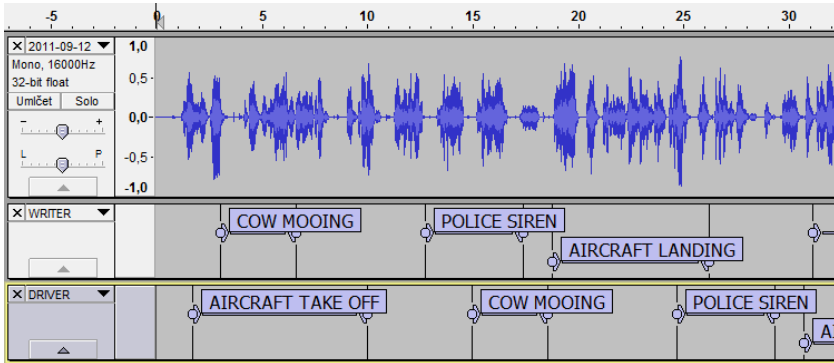


Fig. 1. Noise pattern and alignment of noise tracks for both note-taker and driver

Another lesson learned from the pilot study was about the way how to introduce the noise. Originally we only introduced noise only at the note-taker's side but this did not represent sufficiently the whole variety of natural scenarios. The final tests were conducted with noise added on both sides. Although the same audio was played both to the driver and note-taker, the audio streams were not mutually synchronized so one did not know the noise condition of the other.

Skype communication was also recorded without noise to simplify transcription work needed to analyze the data.

3.2 Dictated Messages

The other important aspect is the correct selection of messages to be dictated. We wanted to have both short and long messages covered in the portfolio of the dictated text due to a suspicion that behavior of the users could differ depending on the message length. One problem was to ensure natural behavior as in normal conditions when the driver dictates messages to a passenger. The method we used was the following. The users were advised to dictate short messages about a selected topic. However the exact formulation was left up to the driver. Regarding the long messages, the drivers were asked to select a topic from their past experience (for example some weekend story) and write the text down prior to the test. The writing exercise ensured that the dictating person does not change the story during the dictation due to some communication problems.

4 Testing Plan

Testing procedure consisted of the following steps:

- Explanation of the test procedure.
- Pre-test questionnaire.
- Learning to use the driving simulator (not recorded).
- First undistracted ride: 3km of LCT driving with no secondary task.
- Dictation of short messages. 2 LCT tests, 3km driving, secondary task is the dictation of semantically prescribed short messages. This part is repeated twice with a different set of messages.
- Dictation of a long message. LCT test, 3km driving, secondary task is the dictation of a previously written long text.
- Second undistracted ride. LCT test, 3km driving, no secondary task.

4.1 The Roles

Each test was conducted under supervision but with minimal intervention of the supervisor with the two participants.

Participants had previous driving experience. To make the most of participants we swapped them so each acted both as the driver and note-taker.

- Driver (D) – drives while performing LCT and dictates message to the note-taker.
- Note-taker (N) – receives the dictated text and records it by typing into a text editor.

5 Evaluation of Results

The test was conducted on 10 individuals. The participants were males – mean age 25.1 (SD 3.8), and they were university students of computer science. The study was conducted in Czech. All participants were native Czech and had Czech cultural background. Before the experiment, each of the participants was instructed verbally on how to proceed with the experiment. The experiment results were anonymized.

By careful analyses of the recorded activities we have identified the following dictation patterns, error correction methods and communication strategies. We state them grouped based on their character. The CHARACTER.NUMBER codes (e.g. C.1, S.1) presented in the next section are references to Fig. 2 (frequency of strategies) and the numbers in parentheses are actual frequencies of each strategy.

5.1 Confirmation Strategies

C.1 Simple acknowledgement (258). When the note-taker is confident, s/he confirms usually only by short confirmation words or sounds (example: “OK”, “Yes”, “Got it”, “Hmm”).

C.2 Acknowledgement by repeating (123). Note-taker confirms what has been understood by repeating the utterance which both confirms that message has been received and provides means to verify the content.

C.3 Request for confirmation (32). A strategy used both by driver and note-taker. The speaker checks if the message has been received (example: “Did you hear me?”).

C.4 Request for confirmation without expected reply (1). It happens that the dictating person asks without real waiting for confirmation (example: “OK?” ...<following dictated text>).

5.2 Synchronization and Communication Channel Handling

S.1 Speeding up or slowing down (30). If the note-taker is not fast enough or dictation is too slow s/he requests slowing down or speeding up by a command (example: “slowly, please!”, or using a confirmation (“I have only got: Buy oranges...”).

S.2 Refinement of the question when no response received (2). If the note-taker does not answer promptly, the driver tends to refine their utterances (example: D: “Barbara” N: long pause D: “Barbara, it’s the English name”).

5.3 Error Prevention

E.1 Ambiguous confirmation (1). Some of the confirmations can be misunderstood and therefore the dictating person adds extra explanation or spelling (example: “Skoda – I mean the car”).

E.2 Active search for quiet intervals (9). The note-taker indicates, on his initiative, the presence of quiet intervals to the driver by short phrases (example: “OK”, “go on”, “now”) which helps to pass the message during favorable noise conditions.

5.4 Error Correction Strategies

R.1 Correction specified by interrogative pronoun (29). The note-taker asks for a specific part which was not understood using an interrogative pronoun question (example: D: “Then you take the mashed potatoes” N: “Take what?”).

R.2 Explanation notes (3). If the note-taker does not understand repeatedly, the driver starts to describe in other words or using explanations (example: “Villon, the French poet”).

Note: We did not see the opposite usage where the note-taker would ask “Do you mean the French poet Villon?” although it is obviously possible.

R.3 Correction specified by context (19). The note-taker asks based on understanding the context (example: D: “We meet at 5 at the church.” N: “am or pm?” D: “afternoon, today afternoon.”).

R.4 Correction specified by order (6). The note-taker asks for a specific part of dictation using item order (example: D: “Buy oranges, apples and butter.” N: “What was the second one?”).

R.5 Correction by repeating a preceding text segment (36). User repeats the whole phrase up to the moment where s/he needs re-dictation (example: N: “Drive towards..?” D: “toward the castle”).

R.6 Correction by repeating a shorter form (1). This is similar to the above repeating of a preceding text segment, but in this case the dictating party only provides a shorter form of the corrected or missing segment, and it is up to the note-taker to finalize the text. Example: D: “Put it to microwave and warm it.” N: (only understood up to “microwave”) “microwave?” D: “warm it”. N: (expands to “and warm it”).

R.7 Spelling (3). Initially, the drivers tended to repeat a problematic phrase two or three times and if this was not understood, they started to spell that phrase. This strategy was used both by the driver and note-taker.

5.5 General Observations

G.1 Shortening dictated segments (2). The participants tended to utter shorter phrases in the noisy environment that what would be natural under silence.

G.2 Not speaking over high noise (N/A). Participants did not like to speak over high noise. Instead they waited or repeated the same segment (example: D: “Please, we will go, we will go, damn some motorcycle is starting here.”).

G.3 Multiple repeats result in complex acknowledgements (4). When multiple repeats are needed, participants tend to fall to more complex acknowledgement schemas (e.g. the above-mentioned acknowledgements by repeating).

G.4 Careful repeating (15). When repeating a phrase, the driver’s pronunciation is more careful, s/he tends to over-articulate.

Note: This is often an issue for man-machine systems as it poses a challenge for ASR models.

G.5 Noise adaptation (N/A). Users adapt to the noise level. What is totally unacceptable at the beginning becomes relatively fine after a period of time.

Note: This is observable even in the evolution of driving distraction statistics as reported in Section 6.1.

G.6 Resignation (4). After some period of repeating, the drivers sometimes resigned and agreed to an approximate version or even to a wrong meaning and proceeded further in dictation.

Note: We did not see the opposite pattern where the note-taker would resign although it is obviously possible.

G.7 Well known topics are easier to handle in noisy conditions (N/A). This holds also for local names, if they are unknown to the user, spelling has to be requested.

6 Discussion

The observations listed above appeared in our experiments with varying frequencies reported below in Fig. 2.

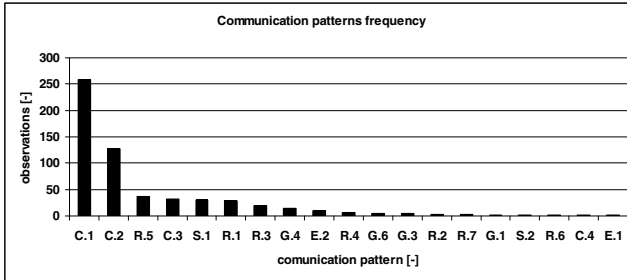


Fig. 2. Frequency of observations in recorded experimental data; see Section 5 for details

6.1 Impact of Dictation on Driver’s Attention

Fig. 3 summarizes distraction caused by the secondary task. Average mean deviation (MDEV) values with standard errors are depicted separately for each task. We adapted the ideal LCT track using the “Undistracted track 2”. The MDEV values for distracted driving were then compared against the “Undistracted track 1” and the difference was analyzed for significance. Participant 1 was excluded from the overall evaluation as an obvious outlier.

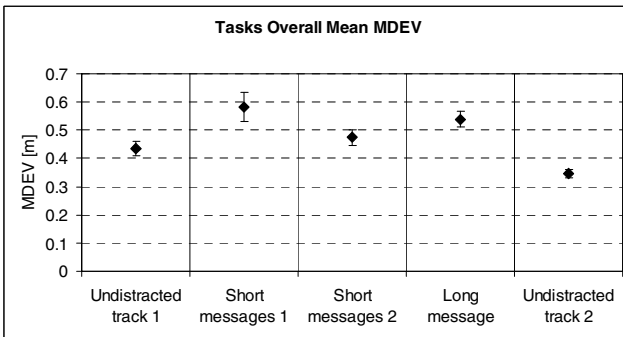


Fig. 3. Average overall mean deviation using data from all participants, except for participant 1 (outlier). Standard error of measurement is displayed.

Fig. 3 shows that driving while dictating under noise was more demanding from the driver’s point of view than undistracted driving (both undistracted tracks). The conventional two-tailed paired T-test for sample means confirmed statistical significance of the difference between mean values of MDEV for “Short messages 1” and “Undistracted track 1” ($p = 0.019$) and between “Long message” and “Undistracted track 2” ($p = 0.013$).

There is a visible difference between the tasks “Short message 1” and “Short message 2” which is caused by a learning effect (see Fig. 3). The difference between “Undistracted track 1” and “Undistracted track 2” is not only due to learning effect as “Undistracted track 2” was used for adaptation of the ideal track.

7 Conclusion

We presented here the results of collecting and analyzing human behavior when dictating both shorter and longer segments of text. Our study was conducted on 10 users, each of which drove 2 undistracted and 3 distracted LCT trips while dictating under noise. In the resulting transcribed data, we collected and described 22 elementary communication patterns and quantified their observed frequencies. The purpose of our work was to gain better understanding of natural usage patterns used by humans in order to utilize them for development of more natural human-computer dictation UIs and dialog systems in general.

As further steps in analyzing human to human communication patterns, we foresee several activities worth pursuing. Firstly the reported tests did not cover some types of distraction including echo or communication channel deterioration with dropping some words. Another interesting area is to proceed by analyzing behavior of the users with visual contact.

References

1. Brostrom, R., Bengtsson, P., Axelsson, J.: Correlation between safety assessments in the driver-car interaction design process. *Applied Ergonomics* 42(4), 575–582 (2011)
2. Cuřín, J., Labský, M., Macek, T., Kleindienst, J., Young, H., Thyme-Gobbel, A., Quast, H., Koenig, L.: Dictating and editing short texts while driving: Distraction and task completion. In: *Proceedings of the AutomotiveUI Conference*. ACM, New York (2011)
3. Mattes, S.: The Lane-Change-Task as a Tool for Driver Distraction Evaluation. In: *Proc. Annual Spring Conference of the GFA/ISOES* (2003)
4. Karat, J., Horn, H., Karat, C.: Overcoming unusability: Developing efficient strategies in speech recognition systems. In: *Proceedings of CHI 2000 Conference*, pp. 141–142. ACM, New York (2000)
5. Kun, A.L., Schmidt, A., Dey, A., Boll, S.: Automotive user interfaces and interactive applications in the car. In: *Personal and Ubiquitous Computing*, pp. 1–2 (2012)
6. Hart, S.G., Stayeland, L.E.: Development of NASA-TLX (task load index): Results of empirical and theoretical research. In: Hancock, P.A., Meshkati, N. (eds.) *Human Mental Workload*. Amsterdam North Holland Press (1988)
7. Pauzie, A.: Evaluation of Driver Mental Workload Facing New In-vehicle Information and Communication technology. In: *IET Intelligent Transport Systems, Special Issue – Selected Papers from HCD* (2008)
8. Yun-Cheng, J., Paek, T.: A Voice Search Approach to Replying to SMS Messages. In: *Proc: INTERSPEECH 2009, 10th Annual Conference of the Intl. Speech Communication Association*, Brighton, United Kingdom (2009)

Multimodal Smart Interactive Presentation System

Hoang-An Le^{1,2}, Khoi-Nguyen C. Mac^{1,2}, Truong-An Pham¹,
Vinh-Tiep Nguyen¹, and Minh-Triet Tran¹

¹ Faculty of Information Technology, University of Science, VNU-HCMC, Vietnam

² John von Neumann Institute, VNU-HCMC, Vietnam

{lhan.ict,mcknguyen.ict}@jvn.edu.vn, ptan@apcs.vn,
{nvtiep,tmtriet}@fit.hcmus.edu.vn

Abstract. The authors propose a system that allows presenters to control presentations in a natural way by their body gestures and vocal commands. Thus a presentation no longer follows strictly a rigid sequential structure but can be delivered in various flexible and content adapted scenarios. Our proposed system fuses three interaction modules: gesture recognition with Kinect 3D skeletal data, key concepts detection by context analysis from natural speech, and small-scaled hand gesture recognition with haptic data from smart phone sensors. Each module can process in realtime with the accuracy of 95.0%, 91.2%, and 90.1% respectively. The system uses events generated from the three modules to trigger pre-defined scenarios in a presentation to enhance the exciting experience for audiences.

Keywords: Smart environment, presentation system, natural interaction, gesture recognition, speech recognition.

1 Introduction

User interfaces aim to provide users with the most convenient ways to use computing systems. To enhance the usability of a system, various approaches have been studied and proposed to mimic natural inter-personal communications: ZeroTouch [11] enhances regular systems with multifinger interaction; a multi-user interaction system [10] allows users to control both desktop and wall-sized environment using depth-sensing techniques like Kinects or Wii remote controllers; a gaze-based interaction system [3] predicts users' behavior based on their looking. With a multimodal approach, Human Computer Interaction (HCI) provides users with not only ergonomic interfaces but also exciting user experiences.

Presentation is a popular activity in daily life such as in lectures, group discussions, or marketing campaigns. However, presenters usually stay away from computers (near the screen or walk around) during their talks, which may interrupt presentations when they go back to their computers for manipulation. Although different kinds of remote controls can be used, it would be more convenient for presenters to deliver their presentations simply by their body motions,

gestures, and speech. This motivates our development of a smart environment that can understand human's behaviors and provide the most inspired and comfortable presentation with high naturalness and flexibility.

Our proposed system fuses three kinds of interaction: users action recognition with Kinect 3D skeletal data, key concepts identification from presenters' natural speech, and hand gesture recognition with smart phones' haptic data. The three kinds of information are captured simultaneously, analyzed in realtime and integrated to trigger certain pre-defined events in a presentation.

The gesture recognition module with 3D skeletal data is based on logistic regression and Dynamic Time Warping and can recognize users' actions with the accuracy of 95.0%. The speech recognition module uses hidden Markov model to recognize speech context automatically with the accuracy of 91.2%. The third module is proposed to overcome difficult situations when sophisticated small hand gestures (with boundary of movement of about 10 to 20 cm) are performed but cannot be recognized with 3D data from a Kinect. The authors use Dynamic Time Warping to process haptic data captured from smart phones' accelerometers to recognize accurately 90.1% of performed gesture.

The content of the paper is as follows. In Section 2, the authors briefly review the development of different methods for interaction with computers. Our proposed system and experimental results are presented in Section 3 and 4 respectively. Conclusions and future work are discussed in Section 5.

2 Related Work

To provide users with ergonomic experience, HCI technologies advance toward intelligent adaptive interfaces with not only unimodal but also multimodal approaches [9]. Each modality is a communication channel that connects the input and the output in an HCI design [6]. In the paper, our proposed system combines three modalities: action recognition using depth data and haptic data, and context recognition based on natural speech data.

There are two main approaches for the action recognition, vision-based or haptic-based, with different difficulties. The vision-based approach depends on the environmental lighting condition while the haptic-based approach requires expensive configuration [9]. The release of Microsoft Kinects with depth sensing technology provides a new trend to solve the difficulties of traditional camera computer vision [16].

For haptic data, it is possible to captured with accelerometer sensors. There are several studies about accelerometer-sensor-built systems for different purposes: Wii controller in presentation system [5], accelerometers in fall detection [15], and culture investigating system [13]. In this paper, dynamic time warping (DTW) is used for gesture recognition (in template matching phase) due to its simplicity, accuracy, and processing speed [12].

Speech recognition systems transform human's spoken speech into digital signal and transcribe the content based on learned data [1]. Since the first application built by Homer Dudley (1930), speech recognition has been developed

and used in different applications: automatically call processing systems, query-based information systems, etc [7, 14]. With the capability to recognize hidden sequences of events, hidden Markov model (HMM) is a typical method in speech recognition [8]. Thus we follow this approach to recognize key concepts in a presentation in realtime from a presenter's natural speech.

3 Proposed System

Our proposed system shown in Figure 1 includes three main phases. In the first phase, the system captures a presenter's gesture data from a Kinect and smart phone sensors, and speech data from a microphone, then dispatches them to appropriate processors in the second phase (c.f. Section 3.1, 3.2, and 3.3). In the last phase, the output gestures and events from the second phase are integrated to trigger corresponding pre-defined scenarios of visualization.

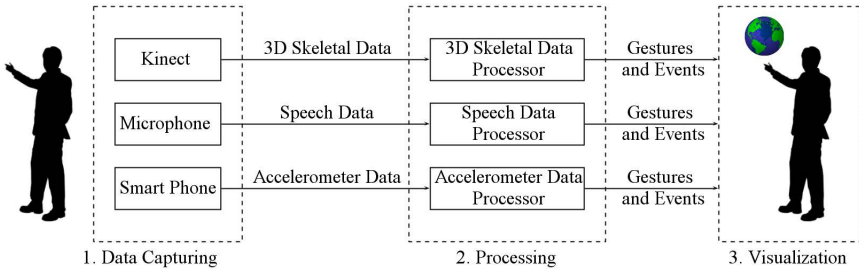


Fig. 1. Overview of the system for the smart presentation environment

3.1 Skeletal Data Based Action Recognition

Figure 2 illustrates 7 typical types of actions that a presenter usually performs during a talk. Although in the current system we focus only on these classes of actions, new categories can be added into the system using the same method to meet users' needs.

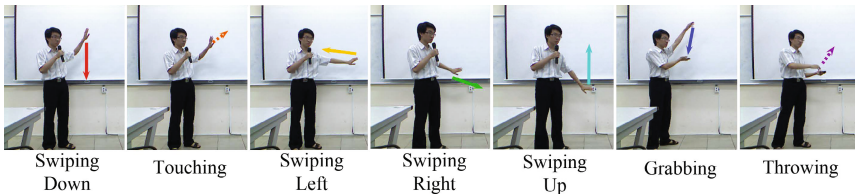


Fig. 2. Proposed action types for action recognition model

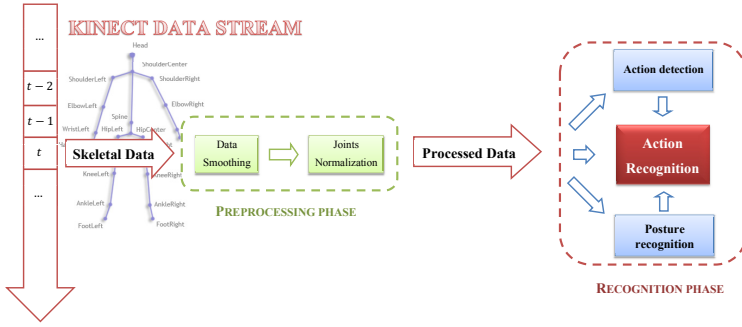


Fig. 3. The presentation action recognition using Kinect skeletal data

The action recognition process consists of 2 main phases (c.f. Figure 3): skeletal data with 20 joints are passed to the *preprocessing* phase, then to the main *recognition* phase.

Preprocessing Phase. The authors use Holt’s double exponential smoothing method to correct the noise by the inference process of Kinect sensors [2]. The joint data is then normalized to be independent from the presenter’s current position and orientation. The spine joint is selected as the origin of the users coordinate, the *x*-axis parallels to the shoulder, the *y*-axis points upward, and the *z*-axis is computed by the cross product of *y*- and *x*-axis.

Recognition Phase. As shown in Figure 4, during a presentation, the movements of a presenter’s hands can be one of the two states: idle (slightly move with respect to the presenter) or active (when he or she is performing an action). By detecting the presenter’s state, the recognition task can be narrowed down to only sequences of active frames. The detection is performed by logistic regression based on the *stability* measurement calculated by the standard deviation of the presenter’s wrists and elbows in a frame sequence.

However, not all the detected active frames belong to meaningful actions but usually initiate special postures. For instance, a swiping down gesture (Figure 2) is usually started by first moving a hand up (frames 1-2) then moving it down (frames 5-7). The learning model used to classify each posture is the logistic regression and one-vs-all technique [4]. The model is trained and tested with different degrees of the feature space to choose the combination that give best prediction.

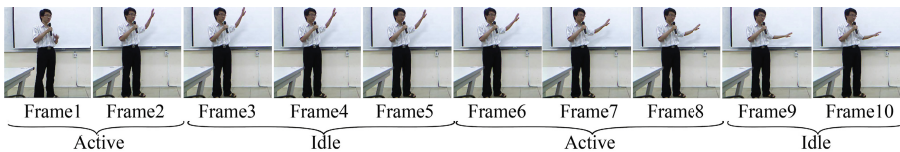


Fig. 4. The sequence of idleness

To recognize patterns of hand movements, features used for classification not only characterize joints' positions but also retain their temporal motion characteristics. The former can be solved by utilizing the information enhanced from the posture recognition subphase; for the latter, the authors construct reference sets consisting of collections of action sequences as a standard to which a given sequence is compared. The similarity between a given sequence and each reference set is used in the feature for recognition. The comparison process between two sequences is done by the Dynamic Time Warping algorithm [12].

3.2 Automatic Speech Recognition

The system records a speaker's commands as spoken speeches then recognizes key terms in the recorded speeches based on the hidden Markov model (HMM). The process has three main stages: data preprocessing, training, and recognition (c.f. Figure 5).

Data Preprocessing Stage. The stage prepares the audio files together with the dictionary and grammar. As Vietnamese language is monosyllabic [14], the phone of a word is kept as the word itself. Although in this paper, the training data cover only Vietnamese digits from zero to nine and solar system's planets, the method can be used with other topics and languages. To increase system performance, users can specify their own set of topics and language.

Training Stage. The training process includes several iterations in which the latter HMM_{i+1} is computed from the former HMM_i . However, a model computed solely on training data are weak to match real life's odds. To increase system's performance, the authors add Gaussian mixtures at every three iterations during the training process. The training process is halted when system performance, which is recomputed after having a new model, does not significantly increase.

Recognition Stage. The new speech signals are transformed into MFCCs before being matched with the training models. The model with highest recognition accuracy is selected. The recognized words, however, might not be the desired ones. Wrong recognition appears because of unlearned words or noisy environment.

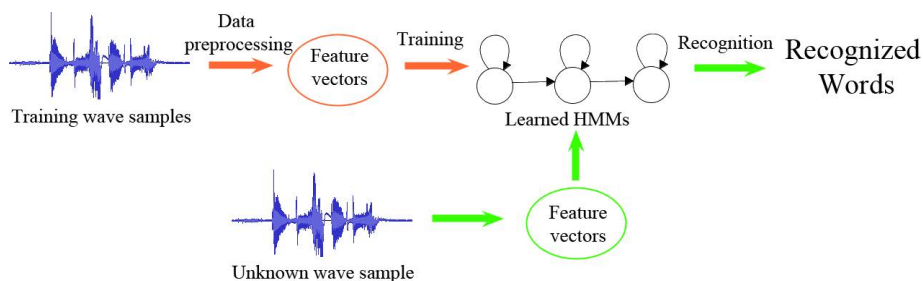


Fig. 5. Overview of the automatic speech recognition module

The authors use a filter to remove noisy words. After the system recognizes the training data, the authors can obtain the distribution of accuracy probabilities and the regular lasting duration of each key term. The filter is a matrix with each row having two thresholds (each is the subtraction of respective mean and standard deviation) of a term’s accuracy and lasting duration. The recognized term is removed if either its accuracy or duration is less than the corresponding threshold.

3.3 Mobile Accelerometer Gesture Recognition

The gesture recognition with data captured from mobile devices’ accelerometers is used to detect delicate gestures that are not appropriate to be recognized with 3D data from a Kinect, e.g. small or occluded hand movements. Accelerometer data is quantized and matched with a template library by dynamic time warping (DTW). Then a gesture can be recognized with the minimum distance (c.f. Figure6).

Data Preprocessing. A sequence of data captured from a mobile device’s accelerometer as a 3D-vector is classified into five classes: moving leftward, rightward, upward, downward, and forward. Because the number of data points collected from a mobile device’s accelerometer varies among samples, linear interpolation is used to sample data at a the sampling rate of 32Hz. Orientation values (tilt, yaw, and roll angles) obtained by a compass sensor are used to normalized the 3D vectors in term of world’s coordinates.

Template Matching. Because DTW can match two series with temporal dynamics [12], it is used in our system to fix nonidentical time interval of gestures without normalizing the data points. DTW is used to calculate the distance between the quantized data and each template in the template library. A sequence is classified into the class of a template with the minimum DTW distance if the distance is less than a certain threshold.

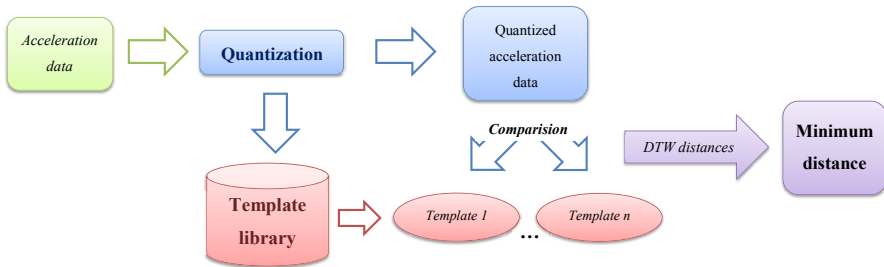


Fig. 6. Overview of the mobile accelerometer gesture recognition module

4 Experiences and Results

4.1 Skeletal Data Based Action Recognition

The data for the action recognition process are collected from 35 clips of 7 action types (c.f. Figure 2). All clips have the same length of 500 frames and the same frame rate of 20fps. To overcome the issue of underfitting and overfitting, the logistic regression models are trained with different degrees (chosen from 1 to 5) and tested with both the training and cross validation sets.

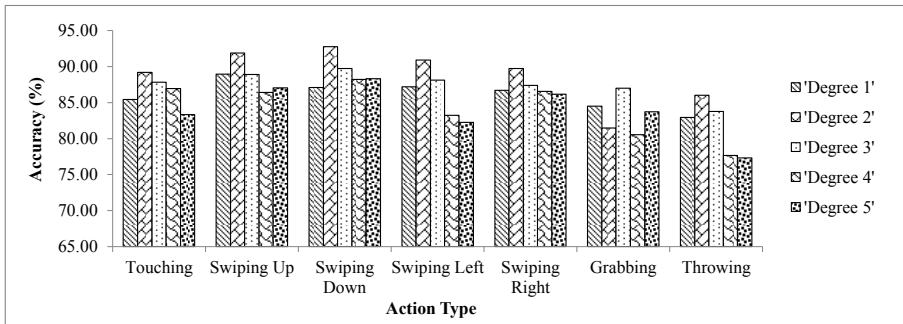


Fig. 7. Recognition accuracy for each action type with different model degrees

As shown in Figure 8, the average accuracy of each action is about 85 – 95%. Although there are some actions, such as swiping up and swiping down, with high accuracy (about 95%), the accuracy of the grabbing and throwing actions are quite low (about 85%). The deficiency of the action's accuracy could be explained as a grabbing or a throwing action (as shown in Figure 2) requires a user to do more rotation to one side in which hand movements are occluded by the user's body and thus can confuse the Kinect skeletal tracking system.

4.2 Automatic Speech Recognition

The training data contains 970 samples of two topics: solar system's planets (550 samples) and digits from zero to nine (420 samples). Data is recorded in wave-form audio format (WAV) with monochannel and sampling rate of 11,025Hz. Figure 8 shows the system's recognition accuracy over different number of Gaussian mixtures and data portions. The experiment suggests that more training samples can give better recognition accuracy as they can cover more real life scenarios. Besides, with a certain data portion, increasing the number of Gaussian mixtures improves the accuracy. The highest accuracy in our experiment is 91.2% and corresponds to the case of 30 mixtures, the highest number of mixtures in experiments.

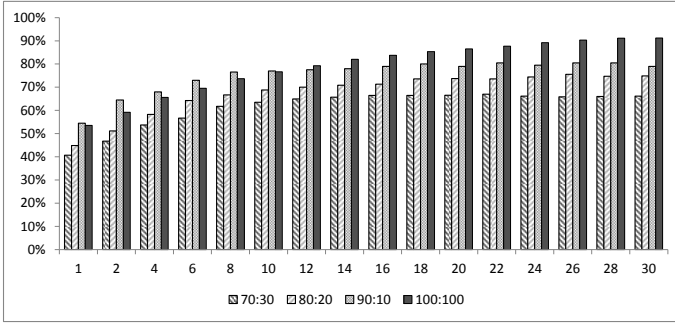


Fig. 8. Recognition accuracy corresponding to different number of Gaussian mixtures (horizontal axis) over four scenarios, each with the ratio between training and testing data respectively be 70:30, 80:20, 90:10, and 100:100.

4.3 Mobile Accelerometer Gesture Recognition

Figure 9 illustrates system’s accuracy using DTW. The data with 200 samples (with the sampling rate of 32Hz) are divided into training and testing sets with the ratio of 1:1. The testing samples are modified so that they are 15 degrees tilted, compared to the training samples. The experiment, however, shows that tilting does not affect much on the system’s accuracy as the average accuracy by DTW can be 90.1%. Besides, data with orientation have higher accuracy than the other case. Therefore, the suitable settings are those with data orientation using DTW method.

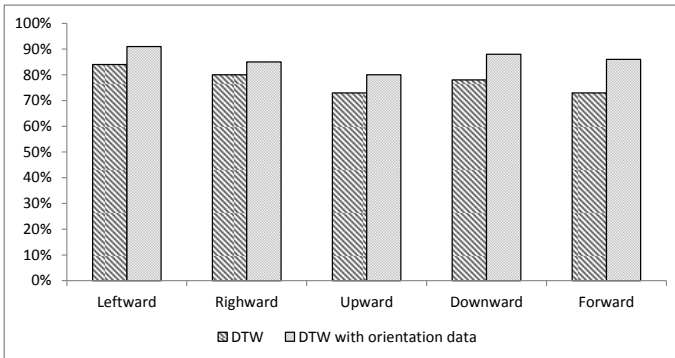


Fig. 9. System’s accuracy using DTW matching method with normal and orientation data

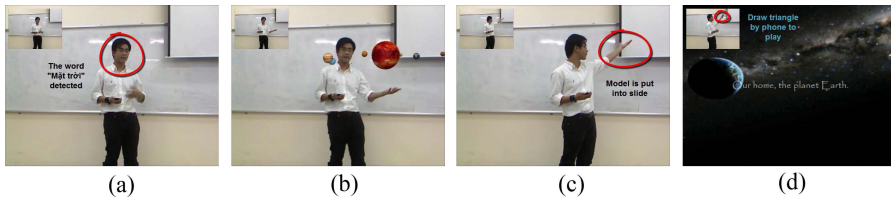


Fig. 10. An experimental scenario of the system: (a) speech recognition, (b) gesture recognition and visualization, (c) putting model into slide, and (d) using mobile accelerometer

4.4 An Experimental Scenario

Figure 10 shows the scenarios originated from a demonstrative presentation. The demonstration window consists of two screens, the small upper left shows the presenter in real life captured by a camera while the rest is what being currently displayed to the audience on the presentation screen. The annotations are interted into the four frames to illustrate triggered events.

Figure 10a indicates the recognition of the term “mat troi” (“the sun” in English) from presenter’s speech, which starts the astronomy lecture. When the lecturer puts his right hand as if he were holding a 3D model of the solar system, the Kinect recognizes the pose and the system displays an augmented 3D model of the solar system above his right hand (Figure 10b). In Figure 10c, the action of moving a hand upward to the presentation screen is recognized as putting the augmented model into the screen and thus opens a full-screen video clip about the solar system. To start the clip, the presenter draws a triangle with a smart phone (Figure 10d), which triggers the “play” event via the phone’s accelerometer.

5 Conclusion and Future Work

This paper introduces a smart interaction system that can react to presenters common and natural behaviors, using their body gestures and spoken speeches. The system has three main modules: gesture recognition using Kinect’s 3D skeletal structure, speech recognition using normal microphones, and hand gesture recognition using mobile devices’ accelerometer.

By experiment, our system can run with high accuracy in real time: 95.0%, 91.2%, and 90.1% for the three modules respectively. Besides, each module can run independently and can be trained with user’s personal data to match demanded reactions, i.e. one can make the presentation of mathematics, physics, or chemistry with different gesture and command set.

For future works, the authors propose to upgrade the system into an interactive framework that allows users to use under several scenarios: meetings, seminars, education, etc. It allows users to integrate different devices (clients) with the framework (server) and provides them with several kinds of interaction and high comfort.

Acknowledgement. This research is supported by research funding from Advanced Program in Computer Science, University of Science and John von Neumann institute, Vietnam National University - Ho Chi Minh City.

References

1. Adams, R.: Sourcebook of automatic identification and data collection. Van Nostrand Reinhold (1990)
2. Azimi, M.: Skeletal Joint Smoothing White Paper (accessed August 13, 2012), <http://msdn.microsoft.com/en-us/library/jj131429.aspx>
3. Bednarik, R., Vrzakova, H., Hradis, M.: What do you want to do next: a novel approach for intent prediction in gaze-based interaction. In: ETRA, pp. 83–90 (2012)
4. Hilbe, J.: Logistic regression models. CRC Press, Boca Raton (2009)
5. Holzinger, A., Softic, S., Stickel, C., Ebner, M., Debevc, M.: Intuitive e-teaching by using combined hci devices: Experiences with wiimote applications. In: Stephanidis, C. (ed.) UAHCI 2009, Part III. LNCS, vol. 5616, pp. 44–52. Springer, Heidelberg (2009)
6. Jaimes, A., Sebe, N.: Multimodal Human-Computer Interaction: A survey. *Computer Vision and Image Understanding* 108(1-2), 116–134 (2007)
7. Juang, B.H., Rabiner, L.R.: Automatic speech recognition - a brief history of the technology development. Elsevier Encyclopedia of Language and Linguistics (2005)
8. Jurafsky, D., Martin, J.: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall Series in Artificial Intelligence. Pearson Prentice Hall (2009)
9. Karray, F., Alemzadeh, M., Saleh, J.A., Arab, M.N.: Human-Computer Interaction: Overview on State of the Art. *International Journal on Smart Sensing and Intelligent Systems* 1(1), 137–159 (2008)
10. Lou, Y., Wu, W., Zhang, H., Zhang, H., Chen, Y.: A multi-user interaction system based on kinect and wii remote. In: ICME Workshops, p. 667 (2012)
11. Moeller, J., Kerne, A.: Zerotouch: an optical multi-touch and free-air interaction architecture. In: CHI, pp. 2165–2174 (2012)
12. Myers, C., Rabiner, L.: A Comparative Study of Several Dynamic Time-Warping Algorithms for Connected-Word Recognition. *The Bell System Technical Journal* 60(7), 1389–1409 (1981)
13. Rehm, M., Bee, N., André, E.: Wave like an egyptian: accelerometer based gesture recognition for culture specific interactions. In: Proceedings of the 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction, BCS-HCI 2008. British Computer Society, vol. 1, pp. 13–22 (2008)
14. Vu, Q., Demuynck, K., Van Compernelle, D.: Vietnamese automatic speech recognition: The fLavor approach. In: Huo, Q., Ma, B., Chng, E.-S., Li, H. (eds.) ISCSLP 2006. LNCS (LNAI), vol. 4274, pp. 464–474. Springer, Heidelberg (2006)
15. Zhang, T., Wang, J., Liu, P., Hou, J.: Fall detection by embedding an accelerometer in cellphone and using kfd algorithm. *International Journal of Computer Science and Network Security* 6(10) (October 2006)
16. Zhang, Z.: Microsoft kinect sensor and its effect. *IEEE MultiMedia* 19(2), 4–10 (2012)

Multimodal Mathematical Expressions Recognition: Case of Speech and Handwriting

Sofiane Medjkoune^{1,2}, Harold Mouchere¹,
Simon Petitrenaud², and Christian Viard-Gaudin¹

¹ LUNAM University, University of Nantes, IRCCyN, France
firstname.lastname@univ-nantes.fr

² LUNAM University, University of Le Mans, LIUM, France
simon.petit-renaud@lium.univ-lemans.fr

Abstract. In this work, we propose to combine two modalities, handwriting and speech, to build a mathematical expression recognition system. Based on two sub-systems which process each modality, we explore various fusion methods to resolve ambiguities which naturally occur independently. The results that are reported on the HAMEX bimodal database show an improvement with respect to a mono-modal based system.

Keywords: Multimodality, graphical languages, data fusion, handwriting, speech.

1 Introduction and Motivation

Speech and handwriting are very common interaction modalities between humans. With the advances of new devices and of robust recognition algorithms it is possible to extend the usage of such input modalities to Human Computer Interaction (HCI) [1, 2]. In this work, from one hand, we are considering online handwriting produced by interfaces like touch-screens, interactive whiteboards or electronic pens. On the other hand, we suppose also available a speech signal which records the corresponding information as uttered by a speaker. In this regard, the same information is supposed to be available but with the intrinsic capabilities and limitations of each of these two modalities. To take advantage of these two information sources, we have experimented a case where naturally each of them conveys differently the processed information knowledge. The case studies proposed in this work concern the problem of Mathematical Expression (ME) recognition. Of course, some existing tools allow entering MEs in a document. Some are very powerful, as the LaTeX language, but they require a high level of expertise. More interactive tools are also available such as the Mathtype equation editor, but, they still suffer from a cumbersome sequence of selections which often delays the ME production. From these observations, it is clear that a more direct way of inputting MEs would be very beneficial. However, this problem is more difficult than text recognition for several reasons. First of all, the mathematical language is composed of a large set of symbols. To cover correctly various domains of sciences, several hundreds of symbols are required. This will

introduce more confusion between symbols. Second and even more important point, the mathematical language is not a one dimensional (1D) language. Indeed, it is not a left-right sequence of symbols, but a two-dimensional layout where the spatial relations play an important role in the meaning of the expression. The extraction of the layout will be even more difficult from the audio signal, since a spoken language is not specifically adapted to put in plain words spatial relationships.

As Fig. 1 shows, speech and handwriting based systems do not have the same drawbacks. Errors committed by each of the two systems may be corrected by the use of the other. So, better performance can be expected by proposing a speech-handwriting system for ME recognition (MER).

The paper is organized in four sections, as follows. In section 2, we describe the global system, by highlighting its main modules. In section 3, we focus on the fusion part. Section 4 is devoted to the experiments: first we check the complementarity of both modalities on isolated mathematical symbols, and then to complete ME. In the last section, we conclude the paper.

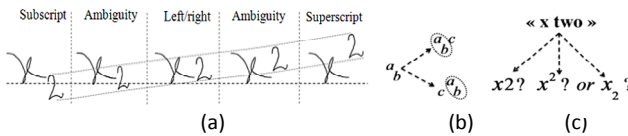


Fig. 1. Some drawbacks that mono-modality based systems encounter, due to the (a) fuzziness nature of the relationships; (b) role of the symbol according to the context; (c) ambiguity of the speech description.

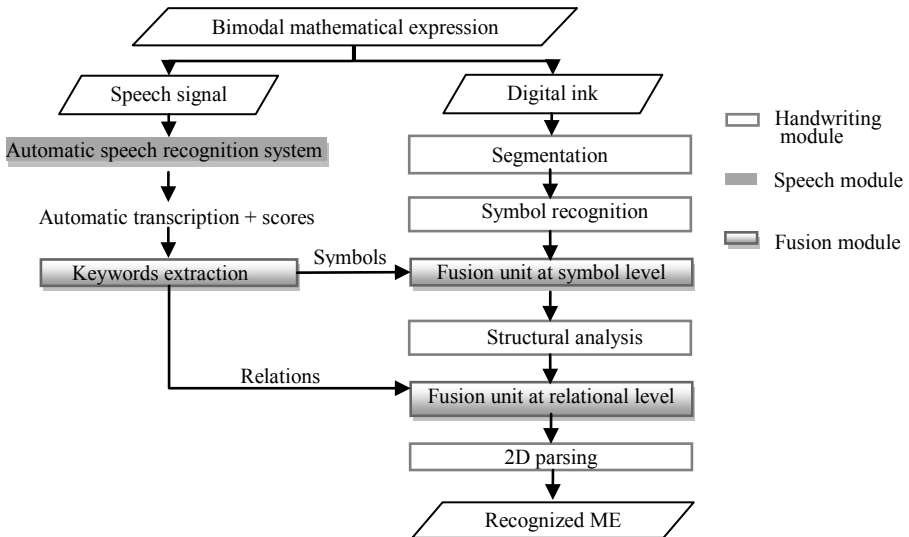


Fig. 2. The collaborative architecture for bimodal mathematical expression recognition

2 Global Overview of the Proposed Method

We propose in this work a combined system composed of two specialized ones: an online handwritten ME system and a speech recognition one. The system in charge of handwritten MER receives as input a set of elementary strokes, and gives a formatted ME as an output. Concerning the system in charge of the audio signal processing, it takes as input the audio signal and provides as a result an automatic transcription which is a textual description of the ME as uttered by the speaker. The information coming from both modalities are merged through the fusion module. This module uses the textual description issued from the speech module and extracts two kinds of information that will be supplied to the handwriting module. The combination process is done using classical data fusion techniques [3] as in Fig. 2. These three modules (handwriting, speech and fusion) will be briefly presented in the remaining of this section. The next section reports a deeper presentation of the main module in this work: the fusion module.

2.1 The Handwriting Recognition Module

The handwriting module has to make the complete interpretation of the handwritten signal and propose the final interpreted version of the ME. This is mainly done at two levels: symbol identification (segmentation and recognition) and relationships discovering (through which the identified symbols are spatially arranged). Thus, recognizing a handwritten ME includes three sequential but interdependent steps [5, 6]: segmentation, symbol recognition and spatial relations interpretation. The aim of the segmentation process is to form the symbol hypotheses " h_s " from the set of strokes. Each " h_s " has to be labeled; this is the role of the recognition stage, where a list of the most probable symbols with confidence scores is assigned to " h_s ". The structural analysis of the global layout including the identified symbol hypotheses is the third step. Finally, the results of these three steps are used to deduce the final ME layout thanks to a bi-dimensional grammatical parsing. Optimizing separately each step has a major drawback since the failure of one step can lead to the failure of the next one. To alleviate this problem, the simultaneous optimization of the segmentation and recognition steps is reported in various works as in [6, 7]. The handwritten MER subsystem used in the architecture of Fig. 2 is largely based on Awal and al.'s system [6].

2.2 The Speech Recognition Module

Using speech for mathematical expression recognition is usually done by means of two successive processes [7, 8]. The first one is a classical automatic speech recognition (ASR) system which provides a textual description of the ME according to the speech describing the ME. The second one is a syntactical-grammatical one. It analyzes the text given by the ASR (1D) to deduce the corresponding ME written in a mathematical language (2D). Thus, even if there are no automatic transcription errors, the relative (un)-clarity of the description might result in ambiguous interpretations. Furthermore, even if both ASR system and speaker are hundred percent accurate, the

bidimensional aspect of the ME is hard to retrieve (*ref.* Fig.1). In the rare existing systems [8, 9], the parsing (1D to 2D) is most of the time assisted by either introducing some dictation rules or using an additional source of information (such as using a mouse to point the position where to place the different elements). This makes the editing process less natural and far from what is expected from this kind of systems.

In our framework, the speech module role is limited to the task of automatic speech transcription providing the textual description. This textual description is then used within the fusion. This ASR task is carried out by a system based on the one developed at the LIUM [10], which is based on the CMU-Sphinx transcription system [11].

2.3 The Fusion Module

The fusion module ensures the connection between the specialized systems (handwriting and speech modules) in order to benefit from the existing complementarity between both modalities. This module is the main contribution of the current work and it is inspired from the data fusion field. Let us first present in the following section this concept and after discuss its use for our purpose: automatic MER.

3 The Fusion Module Description

The idea of multimodal human-machine interaction comes from the observation of the human beings' interaction. Usually, people simultaneously use many communication modes to converse. This makes the conversation less ambiguous. The main goal of this work is to mimic this procedure to be able to set up a multimodal system dedicated to mathematical expressions recognition. Generally, data fusion methods are divided in three main categories [3, 4]: early fusion which happens at features levels; late fusion which concerns the intermediate decisions fusion and the last one is the hybrid fusion which is a mix of the two. Within each approach, three kinds of methods can be used. Rules based approaches represent the first category and include methods using simple operators such as max, (weighted) mean or product. The second category is based on classification techniques and the last one is based on estimation.

In order to accomplish its task (combination of the information coming from both modalities for MER), the fusion module uses the textual description given by the ASR system to assist the handwriting module at two levels: symbol and relation. This why this module can be broken down into three distinct parts: **the keyword extraction unit, the fusion unit at symbol level and finally the fusion unit at relational level.** Since the signals coming from both modalities are heterogeneous and with the objective of using suitable recognition techniques for each modality, we chose to use a late fusion strategy. We give in the following the complete description of each unit.

3.1 Keyword Extraction Unit

The purpose of this unit is to analyze the text describing the ME provided by the ASR system. As a result, two word categories are identified. The first one is composed of

words which are useful for the MER process. They spot either symbols (such as: 'x', 'two', 'parentheses'), either relations ('subscript', 'over') or both ('integral', 'square root'). The second category of words includes all the other words. These words are only used to make sense from the language point of view. Here, we consider the words from the first category as keywords. A dictionary is built in such a way that each symbol and each relation is associated to one or more keywords. For example if the word 'squared' exists in the transcription, the ME under process could contain the symbol '2' and the relation 'superscript'. This dictionary is used within the fusion units to identify the included symbols/relations in the current ME according to speech.

3.2 Fusion Unit at Symbol Level

Besides of considering the late fusion strategy, in this work, we explored some rule based methods to perform the fusion at symbol level. Let us define some notations that we will use to describe the fusion methods we explored. Let us denote by $C = \{c_1 \dots c_N\}$, the set of the N possible symbol classes we consider. If an hypothesis 'x' has to be classified with respect to the modality 'i' ($i \in \{s, h\}$, where 's' represents speech and 'h' is for handwriting), let us define the score of this symbol class 'c_j' assigned to this hypothesis as: $d_{i,j}(x)$. The decision score after fusion is denoted as $d_j(x)$. Now, we focus on the methods used to obtain the scores after fusion.

1. Weighted summation: in this case, the fusion score $d_j(x)$ is given by equation 1.

$$d_j(x) = \sum_{i \in \{s, h\}} w_{i,j} d_{i,j}(x), \quad (1)$$

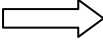
where the $w_{i,j}$ are some weights that can be defined in several ways. These weights can be the same for both modalities (simple **mean**) if we trust in the same way both modalities: $w_{h,j} = w_{s,j} = 0.5, \forall j$. They can depend on the global performances (**meanWGR**), using the global recognition rates R_h and R_s with respect to each modality: $w_{h,j} = R_h / (R_h + R_s)$; $w_{s,j} = R_s / (R_h + R_s), \forall j$. They can also be related to the local performances (**meanWCR**) using the class recognition rates $R_{h,j}$ and $R_{s,j}$ with respect to each modality: $w_{h,j} = R_{h,j} / (R_{h,j} + R_{s,j})$; $w_{s,j} = R_{s,j} / (R_{h,j} + R_{s,j}), \forall j$.

2. Belief functions based fusion (Belief F): the belief functions theory aim's is to determine the belief concerning different propositions from some available information [12, 13]. It is based on two ideas: obtaining degrees of belief for one question from subjective probabilities, and the combination of such degrees of belief when they are based on independent items of evidence. Let Ω be a finite set, called frame of discernment of the experience. The concept of belief function is the representation of the uncertainty. It is defined as a function m from 2^Ω to $[0; 1]$ with $\sum_{A \in \Omega} m(A) = 1$. This quantity $m(A)$ gives the belief that is exactly allowed to the proposition A . Various combination operators are defined in literature. In this work, we focus on the most used and optimal one [13]. It is the Dempster's combination rule. For two belief functions m_1 and m_2 , we obtain \tilde{m} using the conjunctive binary operator:

$$\forall A \in \Omega, \tilde{m}(A) = \sum_{B \cap C = A} m_1(B) m_2(C) \quad (2)$$

In our experiment, the belief functions are deduced from the recognition scores of symbols assigned by the specialized systems. These scores are normalized to be in the range [0, 1]. For example, let us consider H_{hyp} and S_{hyp} respectively a handwriting and speech hypotheses to combine. The recognition processes in both modalities give recognition lists (symbol label and score s). The associated masses (beliefs) can be:

Example of associated beliefs (masses)

<p>Recognized labels list (with scores)</p> <p>for $S_{hyp} = \begin{cases} s(x) = 0.62 \\ s(s) = 0.10 \end{cases}$</p> <p>for $H_{hyp} = \begin{cases} s(n) = 0.52 \\ s(x) = 0.46 \end{cases}$</p>		<p>for $S_{hyp} = \begin{cases} m(x) = 0.62 \\ m(s) = 0.10 \\ m(\Omega) = 0.28 \end{cases}$</p> <p>for $H_{hyp} = \begin{cases} m(n) = 0.52 \\ m(x) = 0.46 \\ m(\Omega) = 0.02 \end{cases}$</p>
---	---	---

The score $d_j(x)$ for a hypothesis 'x' to be the class 'j' is then equals to $\tilde{m}(x)$ obtained from equation 2 in which m_1 represent handwriting masses and m_2 speech masses.

3. Fusion classification based: a support vector machine classifier (SVM) with a Gaussian kernel is used to perform this task. We use the scores from each of the upstream systems as input features of an SVM classifier. Thus, this classifier knows the two score lists provided by each independent specialized classifier ($2 \times N$ features) and computes a new score to every classes (N outputs).

3.3 Fusion Unit at Relational Level

At relational level, the fusion is done during the spatial analysis phase. The parser in charge of this task, in the handwriting modality, explores all the possible relations for each group of elementary symbol hypotheses proposed by the symbol recognition module. For example if we consider the case of two symbols, the relations explored including only these two symbols can be: *left/right*, *superscript*, *subscript*, *above*, *under* and *inside*. For each explored relation a cost is associated [6]. The relation which will be considered in the ME is the one having the smallest cost and satisfying the considered grammar. The fusion at this level is done by exploring the extracted keyword list. If an explored relation exists in this keywords list, its cost is decreased, otherwise it is increased. This is expressed in equation 3, which $RC(R_i)$ and $RC_{new}(R_i)$ are respectively the relational costs before and after fusion for the relation R_i and α_e ($\alpha_e < 1$) and α_p ($\alpha_p > 1$) are respectively parameters to enhance relations present in both modalities and penalize those missing in the speech modality:

$$RC_{new}(R_i) = \begin{cases} \alpha_e RC(R_i) & \text{if the relation } R_i \text{ is in the keyword list} \\ \alpha_p RC(R_i) & \text{otherwise} \end{cases} \quad (3)$$

4 Results and Discussions

In this section we present two kinds of results. The first one is to validate the hypothesis of the existing complementarity between speech and handwriting modalities.

We propose a first experiment considering only the recognition of isolated symbols. In this case, in both modalities, the symbols are already segmented and no relations between symbols exist. Thus, this experiment is a very simplified scenario which does not completely match with the real-life application. Then a more complete experiment is presented including all the steps of the MER process.

4.1 Database Description

The data used to perform the experiment is from the HAMEX [14] database. This database includes a set of approximately 4350 ME, each of them available in the spoken and the handwritten modalities. The vocabulary covered by HAMEX contains 74 mathematical symbols, including all the Latin alphabet letters, the ten digits, six letters from the Greek alphabet and various mathematical symbols (integral, summation...).

4.2 Case of Isolated Mathematical Symbol (Gain at Symbol Recognition Level)

The on-line handwriting recognition is performed by the symbol recognizer used in the global MER system of Fig.2. It is globally based on a Time Delay Neural Network (TDNN) classifier [6]. The corresponding output is a list of Nbest classes with their scores which are normalized in the range [0, 1]. The isolated spoken words recognition is performed using a system based on MFCC coefficients and template matching using a DTW algorithm [15]. Here again, a list of most probable symbols with scores in the range [0, 1] is given.

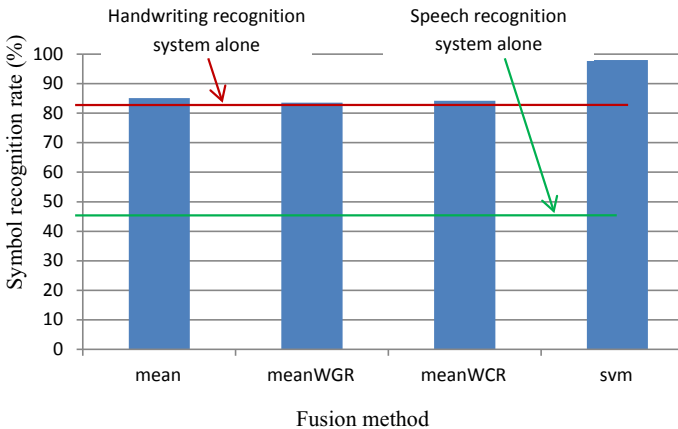


Fig. 3. Recognition rates before and after fusion at symbol recognition level

As we can see on Fig.3, the bimodal based recognition outperforms the mono-modality based systems regardless of the used fusion method. The classification based approach appears to be the best fusion method (recognition rate of 98.04% against the highest recognition rate in the mono modality mode, 81.55%). This classifier takes

clearly advantage of the strengths and weaknesses of each individual classifier because of its training stage, while other combination methods are simpler, they rely on more heuristic functions.

4.3 Case of Complete Mathematical Expression

In the case of a complete ME, the architecture in Fig.2 is used. The fusion process, here, is more complicated (cf. section 3.3). The handwriting recognition task is accomplished with the online handwritten MER system we participated with for CROHME2012¹ competition [16]. A set of 500 ME from the HAMEX train part is used to tune the fusion parameters (cf. equations 1, 2 and 3). The results reported here concern a set of 519 ME of the HAMEX test part selected in such a way to satisfy to the CROHME grammar (task 2). Finally, the models of the ASR system are trained on the whole speech data of the HAMEX train part. Concerning the fusion process itself, the selection of the speech segment to combine with the handwriting group of strokes is done according to the labels intersection in the top N (N is set experimentally to 3). Thus a handwriting segmentation hypothesis is combined with a speech segmentation hypothesis only if a common label in the 3best recognition lists from both modalities exists. Since at the moment of running this first experiment the alignment at the ground truth level of the handwriting and speech streams is still not available, the classification based fusion is not explored. We report on Fig.4, the recognition rates at the expression level for the different fusion methods explored.

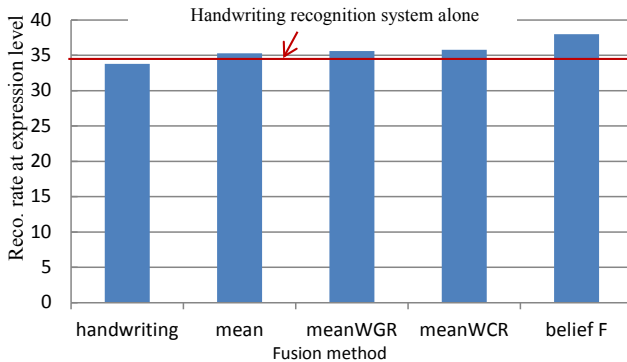


Fig. 4. Recognition rates at expression level before and after fusion

Similarly to the case of isolated symbols recognition, the fusion process improves the performances compared to a purely mono-modality based system. The recognition rates at expression level show that whatever the fusion strategy used, the performance is better. The best fusion configuration is the one based on the belief function theory. The exploration of the various mean weighted methods showed that a good weighting of the scores coming from both modalities is important, since it allows dealing with

¹ <http://www.isical.ac.in/~crohme/>

the problem of score normalization in both modalities. This support the hypothesis that using a classification based approach can fix the score normalization problem.

Deeper analysis comparing the best fusion and the handwriting systems, reported in table 1, show the fusion gain at lower levels (segmentation and recognition).

Table 1. Performances comparison of handwriting and belief functions fusion based systems

Evaluation level in [%]	Stroke classification rate	Symbol classification rate	expressions recognition rate with		
			exact match	1 error at most	2 errors at most
handwriting system	80.05	82.93	34.10	46.44	49.52
fusion based system	83.40	85.40	38.34	50.10	53.37

The improvement brought by the fusion process concerns both low (strokes and symbols) and high (complete expression) levels. Another important remark is that when allowing only one error (symbol or relation), we gain around 30% of ME (from 38.34% to 50.10%); this suggests that there is still scope for additional contribution of the fusion process, especially by exploring classification fusion methods. We give in Fig.5 a real example of results, where the handwriting system fails to provide the right solution when the fusion one, thanks to this bimodal processing, succeeds on this task.

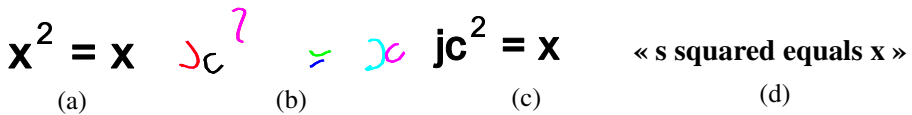


Fig. 5. Real example of a contribution of the bimodal processing (misrecognized in handwriting and recognized in fusion); (a) ME ground-truth, (b) its handwritten version, (c) the recognized result without fusion, (d) the automatic transcription of its spoken description

In this example, the first two strokes (going from the left, in Fig.5-b) should belong to the same symbol. However during the handwritten recognition, combining both of these strokes into the same symbol hypothesis leads to its misclassification. Indeed, the classifier suggests that this segmentation is not valid and assign a high score for rejection label 0.84 and answers, in a second rank, that it can be an 'x' with a score of 0.15. When fusing, this segmentation hypothesis is combined with the audio segment containing also an 'x' label as a recognition hypothesis. Unfortunately, apart from the belief functions fusion method, all the other methods do not allow to recover the right label. This is mainly due to the fact that in the audio segment also, there is a conflict between the classes 's' (0.48) and 'x' (0.45). The belief functions method, by modeling a part of ignorance (equation 2), makes the 'x' label score high enough to rank it as a first hypothesis and to include it during the structural analysis process.

5 Conclusion and Future Work

After a first experiment on isolated symbols recognition to prove the existing complementarity between speech and handwriting, we proposed a new architecture for complete MER based on bimodal processing. The obtained results are quiet satisfying since the performances are improved compared to a mono-modal system.

In a future work, we plan to improve the choice of the couple (speech hypothesis segment, handwriting hypothesis group) to be fused, by exploiting the temporal information in both modalities. The final goal is to reach the best possible synchronization between the two streams. Another interesting point to explore is the use of word lattice from the ASR system, which can provide more information for a considered speech segment. Beside of that, the context of the symbol or the relation is still not used. We believe that this can improve also the accuracy of the global system.

Acknowledgments. The authors would like to thank the French Region Pays de la Loire for funding this work under the DEPART project <http://www.projet-depart.org/>.

References

1. Karray, F., Alemzadeh, M., Saleh, J.A.: Human-Computer Interaction: Overview on State of the Art. *IJSSIS*, 137–159 (2008)
2. Jaimes, L., Sebe, N.: Multimodal human computer interaction: A survey. *Computer Vision and Image Understanding* 108, 116–134 (2007)
3. Thiran, J.-P., Marquès, F., Bourlard, H.: *Multimodal Signal Processing - Theory and Applications for Human-Computer Interaction*. Elsevier (2010)
4. Atrey, P.K., Hossain, M., AEl Saddik, A., Kankanhalli, M.S.: Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 345–379 (2010)
5. Zanibbi, R., Blostein, D.: Recognition and retrieval of mathematical expressions. *IJDAR* 15, 331–357 (2012)
6. Awal, A.-M., Mouchère, H., Viard-Gaudin, C.: A global learning approach for an online handwritten mathematical expression recognition system. In: *PRL*, pp. 1046–1050 (2012)
7. Rhee, T.H., Kim, J.H.: Robust recognition of handwritten mathematical expressions using search-based structure analysis. In: *ICFHR*, pp. 19–24 (2008)
8. Fateman, R.: How can we speak math? University of California, Tech. report (2012)
9. Wigmore, A., Hunter, G., Pflugel, E., Denholm-Price, J., Binelli, V.: Using automatic speech recognition to dictate mathematical expressions: The development of the talkmaths application at Kingston University. *JCMST* 28, 177–189 (2009)
10. Deléglise, P., Estève, Y., Meignier, S., Merlin, T.: Improvements to the LIUM French ASR system based on CMU Sphinx: what helps to significantly reduce the word error rate? *Interspeech* (2009)
11. Cmu sphinx system,
<http://cmusphinx.sourceforge.net/html/cmusphinx.php>
12. Denooux, T.: Conjunctive and disjunctive combination of belief functions induced by non distinct bodies of evidence. *AI* 172, 234–264 (2007)
13. Smets, P., Kennes, R.: The transferable belief model. *AI* 66, 191–234 (1994)
14. Quiniou, S., Mouchère, H., Peña Saldarriaga, S., Viard-Gaudin, C., Morin, E., Petitrenaud, S., Medjkoune, S.: HAMEX – a Handwritten and Audio Dataset of Mathematical Expressions. In: *ICDAR*, pp. 452–456 (2011)
15. Muda, L., Begam, M., Elamvazuthi, I.: Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient and Dynamic Time Warping. *TJC* 2 (2010)
16. Mouchère, H., Viard-Gaudin, C., Kim, D.H., Kim, J.H., Garain, U.: *ICFHR2012: Competition on recognition of online handwritten mathematical expressions (crohme 2012)*. In: *ICFHR* (2012)

'Realness' in Chatbots: Establishing Quantifiable Criteria

Kellie Morrissey and Jurek Kirakowski

School of Applied Psychology, University College Cork, Ireland
{k.morrissey, jzk}@ucc.ie

Abstract. The aim of this research is to generate measurable evaluation criteria acceptable to chatbot users. Results of two studies are summarised. In the first, fourteen participants were asked to do a critical incident analysis of their transcriptions with an ELIZA-type chatbot. Results were content analysed, and yielded seven overall themes. In the second, these themes were made into statements of an attitude-like nature, and 20 participants chatted with five winning entrants in the 2011 Chatterbox Challenge and five which failed to place. Latent variable analysis reduced the themes to four, resulting in four subscales with strong reliability which discriminated well between the two categories of chatbots. Content analysis of freeform comments led to a proposal of four dimensions along which people judge the naturalness of a conversation with chatbots.

Keywords: Chatbot, user-agent, intelligent assistant, naturalness, convincing, usability, evaluation, quantitative, questionnaire, Turing, Chatterbox.

1 Evaluating for Naturalness

Conversational agents, or chatbots, are systems that are capable of performing actions on behalf of computer users; in essence, reducing the cognitive workload on users engaging with computer systems. There are two key strategies used. The first is the use of a set of well-learned communicative conventions: natural language and the accepted conventional structure of a conversation so that the user does not need to learn artificial conventions (such as SQL, other query languages, or highly constrained programming methods.) The second is enabling the user and the computer to refer to broad shared classes of knowledge of which either the computer, the user, or both hold the specific details so that the solution to a problem can be arrived at by negotiating through the knowledge space in a way that neither side need concern themselves with details which are difficult or impossible for that side to represent.

Implementation of these two strategies: naturalness of interaction and sharing knowledge space are the two essential features of all conversational agents. Different agents vary in the success of their implementations of each. But the important point is that as far as the user is concerned, the interface is one: an intelligent conversation heeds conventional structure as well as being about a shared referent.

Chatbots are over fifty years old. Turing [19] in his famous thought experiment set up what is considered to be the touchstone of evaluation, despite some researchers'

claims (e.g., [10],[11]) that Turing's test offers neither an operational definition of intelligence in computers nor necessary and sufficient conditions for the demonstration of such intelligence. Weizenbaum in 1966 [20] wrote his ELIZA in the MAD-SLIP programming language and demonstrated how one could simulate human conversation through simple pattern-matching of user input to the data stored in its script. Weizenbaum's complaint that ELIZA was often not seen as the trick it really was and his vision that a more successful ELIZA would build a model of the person with whom it converses during the conversation is perhaps a symptom of the representation-dominated theories of the mind current at the time, embodied most starkly in Fodor's landmark book *The Language of Thought* [9].

There is a potentially wide range of applications for chatbots today. These types of agents have a part to play in domains involving the negotiation of information retrieval and organization. As the amount of knowledge held in data stores expands, and as the technical level of skill required by the average user diminishes to make this knowledge available to an increasingly diverse user population, intelligent agents become increasingly important to the universal acceptance of technology. Chatbots are found in stores and help sites as embedded online assistants, in chatrooms as spam agents, and in video games as non-playing characters. Notable applications of recent chatbot-like technology in the press are IBM's Jeopardy-winning computer Watson [5] and Apple's embedded "personal assistant", Siri [1].

So what makes for a convincing, satisfying, perhaps a natural interface for a user agent?

It is perhaps to answer this question that challenges such as the Loebner Prize still run Turing-like tests each year in an attempt to spur on the creation of a chatbot that can converse in a naturalistic fashion. The Loebner Prize was started by Dr Hugh Loebner in 1991 and has been held every year since then, with multiple entrants each year. While the prize money of \$100,000 which has been set aside for the winning chatbot undoubtedly inspires many programmers to create and improve their chatbots, the Loebner Prize has been criticised for a lack of realism. Shieber [18], in attendance at the first Loebner Prize, contends that the Loebner Prize is not a true representation of Turing's test: the conditions of the challenge are modified extensively in order to allow the chatbot what is considered to be a fighting chance – the topic of the human-computer conversation itself is restricted to a singular domain and is thus not a free test. The binary "yes/no" decision of the original Turing test, Shieber observed, is also replaced by a ranking format, in which the judges rank in order of their "humanness" but not specifying an absolute "human" threshold. Is there really a point in running a test in which such large concessions need to be made?

Cohen, in a paper entitled "If not the Turing test, then what?" [7], describes a number of differential intelligence tests for bots. Cohen suggests such trials be drawn from the sort of tasks that third graders in a North American elementary school ought to be able to complete. When considering the sort of tasks that Cohen suggests, it is interesting to note that the majority of these tasks require a significant ability to understand, manipulate and produce concepts behind language. While this kind of ability is considered important to demonstrate intelligence in humans since the days

of Binet's pioneering demonstrations [3] it is doubtful whether this is even relevant to the evaluation of naturalness in chatbots.

However, Cohen's *criteria for a good challenge* are fully in accord with the aims of the research presented here. They are threefold:

1. such a challenge must produce feedback that is more developed than the non-gradated "yes/no" of the Turing test,
2. it must have a sense of monotonicity, allowing for repeated reproductions of the challenge to verify the results of previous challenges, and
3. it must "capture the hearts and minds of the research community" - while the Loebner Prize and the Turing test have certainly engendered a large amount of discussion, very few working in the area of intelligent systems would seriously put Cohen's *specific tasks* forward as a measure of the results of chatbot development.

Shawar and Atwell [17] propose "glass box" and "black box" methodologies in order to assess a chatbot. These methodologies represent two sides of appraisal: glass box methodologies assess a given conversation technically for grammar, syntax, sentence structure and appropriateness of answers; while "black box" methodologies broadly attempt to measure user satisfaction. In testing these methodologies using a goal-based task and an Afrikaans-literate chatbot, Shawar and Atwell found that such a separation was ill-suited to the task at hand and proposed that the Loebner Prize criterion of naturalness is in the end perhaps preferable. Their final suggestion is that chatbot success should be functionally defined: "the best evaluation is based on whether it achieves that service or task." In general, we would agree with such a task-based criterion. But chatbots are no longer predominantly used for work-based tasks in the sense of the ISO 9241 part 11 definition of usability [13]. So what to do then, when the chatbot may be designed for nothing more than to be a partner in an amusing natural-seeming human conversation?

Semeraro et al. [16] used a top-down approach to evaluate their agent-based interface, constructing a questionnaire which assesses the chatbot's ability to learn and to aid the user, its comprehension skills, ease of navigation, effectiveness, impression and command. Hung, Elvir, Gonzalez and DeMara [12] note that this is a subjective approach: a criticism which bolsters the need for a statistically reliable evaluatory instrument. They also note that it is more of a general indicator of performance, rather than an appraisal which would lead to generalisable findings for chatbots. Rzepka, Ge & Araki [15] use a similar 1-10 rating system assessing naturalness and technical ability to continue a conversation in assessing the performance of older-style ELIZA chatbots and newer commonsense retrieval bots, which was then expressed as a "naturalness" degree and a "will of continuing conversation" degree. The issue with these methodologies, however, is that the scales and questionnaires used to test the chatbots are not themselves verified as sufficient by ordinary users and lack reliability and validity as measuring instruments.

The research question addressed in this paper is part on an ongoing research programme to generate measurable criteria for the naturalness of chatbot dialogue that are acceptable to people who are more interested in the results of chatbot development than the technical issues of the development itself. Although at this stage we feel we

can make a modest contribution to our knowledge on the subject we still have a way to go, as we will explain in the conclusion to this paper. There are two studies which we wish to present.

2 Study 1: In the Words of the Users

This study used a chatbot that was modelled on the ELIZA scheme but which explicitly made use of situational semantics [2]. Information, according to this view, exists in situations - which are usually local and most probably incomplete. Users, who in this case are considered to be the environment within which the chatbot finds itself, could be considered to have a large amount of information that is part of the situation and which therefore does not need to be represented in the software databases. Conversations can be created according to topics, in which there may be a number of types, and the contents of conversations are ELIZA-type input-output transformations, which are considered as tokens linking to the types in the conversation topic. The program worked on a simple subsumption architecture [4], in that there are three layers, each of which could hold the floor at any one moment and which communicate with the other layers by very simple excitatory signals.

- Layer 1: Conversational maintenance on a given topic where tokens are connected to types and types are connected to other types;
- Layer 2: A switching agent to find a new topic and connect to it;
- Layer 3: General purpose social control: phatic conversational tokens.

The program very explicitly did not store previously unused keywords to pop back if it got stuck, the way ELIZA did: instead firstly social control sought to bring the conversation back to track, and then after a while the switching agent came in to negotiate a new topic with the user. Many tokens were created by the developers of each type, and tokens were selected by the program on a pseudo-random basis from each type when required. A randomly-generated time interval preceded each response by the chatbot. The chatbot was given the name of “Sam” with no particular acronym in mind.

In the experiment fourteen participants were asked to interact with the chatbot as described above for three minutes and then to participate in the elicitation of critical incidents with a transcript of their session. Participants were all tested individually in a HCI lab with one experimenter present. No participant was under the illusion that they were communicating with anything other than a chatbot after a few exchanges, although no explicit cues were given by the experimenters. Respondents were tested in the vicinity of a half-silvered mirror behind which a dialogue partner might have sat.

The Critical Incident Technique [8] requests the respondent to identify particular moments in an experience that the respondent, in hindsight, considers to have been critical during the experience. At the end of the interaction, therefore, the participants were presented with a printed transcript of the dialogue and asked to highlight instances of the conversation that seemed particularly unnatural (up to three examples)

and then were asked to explain why this was so. The same was done for up to three examples of the dialogue that did seem convincing. By the time the respondents had marked the transcript up and been interviewed, it was very clear to each respondent what they had participated in.

The data produced by the critical incident technique were analysed by content. Three raters participated overall. No particular brand of qualitative analysis was considered to be specifically appropriate, although Grounded Theory [6] might come closest. The first rater went through the user responses and identified each response as belonging to one specific theme to do with having a conversation. No themes were created *a priori*, they emerged as a best fit from the data. The data coding was cross-checked independently by a second rater. Inter-rater reliability of approx. 0.53 was obtained in the first pass, which is low (but not usually assessed in Grounded Theory approaches anyway.) Items on which there was disagreement were discussed and placed in mutually agreeable categories with the moderation of a third independent rater. The researchers were reasonably sure in the end that the categories that emerged represented reproducible aspects of the data set.

In general, the kinds of comments reflected the strengths and weaknesses of the three-layer architecture of the chatbot as implemented, and also showed that respondents in general thought that both communicative conventions and the shared knowledge space were of concern when considering the naturalness of the conversation.

A reassuring symmetry emerges in the themes identified by users. For instance, being convincing or not: maintaining a theme is convincing, while failure to do so is unconvincing; colloquial or conversational English is convincing while formal or unusual language is the opposite. Reacting appropriately to a cue is human while failing to react to one isn't. Delivering an unexpected phrase at an inappropriate time does not impress, but damage control statements can rectify the situation. This research was reported by Kirakowski, O'Donnell and Yiu in 2009 [14] who give a full account of each of the seven themes extracted. They are, in summary:

1. Maintenance of themes
2. Responding to a specific question
3. Responding to social cues
4. Using appropriate linguistic register
5. Greetings and personality
6. Giving conversational cues
7. Inappropriate utterances and damage control.

However, there is no indication as to the perceived relative severity of failures by the chatbot. In other words, it is difficult to tell if users found the chatbot's inability to maintain a conversational theme to be a more serious problem than the delivery of inappropriate utterances during the dialogue, or even if there is a degree of individual difference involved in which characteristics of the chatbot's linguistic register are pertinent to its seeming to be natural.

3 Study 2: Towards Quantification

This study developed the first draft of a questionnaire. Questionnaire statements (or items) were created following the structure of the seven themes in Kirakowski, O'Donnell & Yiu's paper ([14], henceforth called the KO'DY structure.) For each theme, at least 4 statements were initially generated that attempted to capture the substance of the theme. Two preliminary validation steps were carried out.

Firstly, face validity of the items was assessed in a meeting of a research group attended by 12 experienced researchers and postgraduate students in the field of cyberpsychology, many of whom also had psychometric expertise. Secondly, the pool of items was then reassessed according to the HCI literature by following articles published relevant to the keywords in the items, in order to increase construct validity. The final inventory consisted of 23 items, with 2-4 statements attempting to measure each KO'DY theme. Items were randomised so as to avoid order effects.

The answering format was a five point frequency scale with the anchors "always", "often", "sometimes", "seldom" and "never" - the rationale for this five point scale was that it matched the statement format of the items. An open-ended question at the end of each evaluation form asked the participant "what do you think this chatbot is best at?"

Participants were chosen from the undergraduate population of University College Cork, Ireland, and were 11 male and 9 female between the ages of 19 and 30 with a mean age of 23. They were all fully briefed as to the nature of the experiment.

Chatbots were chosen on the basis of their placing in The Chatterbox Challenge; an annual chatbot competition along the lines of the Loebner Prize. Five winning entrants were chosen from the 2011 competition to act as the "good" chatbots and five entrants which failed to place in the 2011 competition acted as the "poor" chatbots. As multiple independent judges assess these bots in the Chatterbox Challenge, test-retest reliability should be good, as should, one hopes, be the case for objectivity. Each participant had to evaluate all ten chatbots. The chatbots were presented in a Latin Square design to minimise order and sequence effects.

The apparatus used was a Dell computer running Windows 7 and Google Chrome connected by fast ISDN to the Internet. A basic HTML interface was created for the purposes of the study, briefly listing instructions for the participant and containing internet links to the ten chatbots the participant was to encounter. Clicking on each link in turn opened a new tab which then loaded the page in which the chatbot was embedded. Participants chatted with each chatbot for five minutes on a topic of their choice. After five minutes they filled out an evaluation questionnaire for the chatbot and went on to the next.

Although it would have been straightforward to compute an overall score for each chatbot by summing the 23 items, computing reliabilities, and carrying out an analysis of variance, we were tempted to go slightly further and to attempt to find out how the matrix of 10 x 20 x 27 data items factorised, expecting to find a factor structure similar to the KO'DY structure. We are aware that because each respondent is each responsible for 10 questionnaires within the dataset there may be an amount of spurious intercorrelation between the chatbot scores which is impossible to estimate - but

which might make the matrix more difficult to solve because of multicollinearity. We thus present these results with some reservations.

We put the data through a Principal Components Analysis (PCA, using the SPSS 18 package): PCA highlights the existence of linear components in the data set and assesses the percentage variance that these components contribute to the overall variance. This serves to narrow the scope of the statistical analysis. The contribution of eigenvalues, scree plot and item interpretation allows for an initial data reduction at this stage of statistical analysis. A varimax rotation was then utilised. This transformation of factor loadings allowed for a clarified interpretation of the results.

The initial correlation matrix contained many coefficients of .4 and above. The Kaiser-Meyer-Olkin measure indicated the sampling adequacy for the analysis, $KMO = .92$, and Bartlett's Test of Sphericity achieved statistical significance, 2303.853, $p < .001$, verifying that correlations between items were sufficiently large to justify PCA. The best solution which made sense was with four factors, this explained 59.86% of the variance (we tested from two up to eight factors but the scree plot of eigenvalues showed a definite bend after four and the semantics of the rotated factors supported this.) Varimax rotation on the four factors showed a clear, simple structure. We further conducted a Cronbach's alpha coefficient for reliability and found all four scales achieved 0.70 or higher, thus satisfying the lower level of reliability criteria for scales suitable for research purposes.

The factors are as follows:

1. Factor one, broadly labelled **Conscientiousness** is the largest factor, comprising of ten items which measure how the chatbot seems to keep track of the conversation at hand and how appropriate its responses were. Conscientiousness had a high Cronbach's Alpha of 0.915.
2. The second factor was labelled **Manners**, consisted of 6 items and assessed the ability of chatbots to display polite behaviour and conversational habits. Greetings, apologies, social niceties and introductions were constructs measured in the items within this factor. The Cronbach's alpha coefficient for the factor was .763.
3. The third factor was labelled **Thoroughness** and consisted of 4 items, measuring the formal grammatical and syntactical abilities of the chatbot. The Cronbach's alpha coefficient for the factor was .726. This factor had a large effect size (Cohen's $d = 1.2$) when we came to analyse differences between chatbots (see below), which suggests that the figure for alpha is depressed simply because there are only 4 items in the factor and that this is an important factor.
4. The fourth and final factor was **Originality**, consisting of three items which measured the chatbot's ability to produce what seemed to be original material and also its ability to take the initiative in conversations. The Cronbach's alpha coefficient for the factor was .735 and it is most probably also affected by the small item size.

Given that these four extracted factors seem to have good reliabilities, we then conducted a $2 \times (5) \times 4$ way analysis of variance to establish whether the overall questionnaire was able to distinguish between good and poor chatbots, and whether there were any differences in the profiles of the average good and average poor chatbot. There was a significant main effect of quality of chatbot, meaning that overall the

questionnaire does discriminate well ($p < 0.01$). There was also an interaction between quality and scales ($p < 0.01$) in which Thoroughness gave rise to the biggest difference: in other words, Thoroughness is the biggest discriminator between good and poor chatbots. Manners gave rise to the smallest difference, although it was still statistically significant ($p < 0.01$) as were the differences on all the factors.

The open-ended questions “what do you think this chatbot is best at?” were coded for content analysis; again, tending towards a Grounded Theoretic approach [6]. Codes adhered closely to the data and attempted to explain, conceptually, what was occurring in each line of the participants' answer. In all, four salient themes were present in the data.

Asking and Answering: The transactional nature of the conversation taking place between user and chatbot is highlighted here.

Originality: Users often point out the originality (or lack thereof) of some chatbot responses.

Personality: Participants refer often to the chatbot as having a “personality”: the issue of manners and politeness in chatbot discourse is one which arises time and again.

Relationship with User: An interesting theme which may point towards somewhat of a sense of intersubjectivity between chatbot and user – however illusory that intersubjectivity may be!

4 Combining the Results and the Next Step

A picture of what non-technical users are expecting from a chatbot is beginning to emerge, although the final step is to revise the current 23 items by adding items suggested from the content analysis of the second study to balance up the scales, revising those items from the first study which loaded less well on the four original factors, and then conducting an exploratory followed by a separate sample confirmatory study. If all goes well, we should, by the end of the confirmatory study phase be able to offer a relatively short instrument with high reliability, validity, and a reference database against which we can score the percentile of naturalness at which a chatbot performs. At present we are not able to do this, and we reserve not to publish the current item bank and its loadings on the grounds that it is work in progress.

However, as a summary, we can see four broad dimensions on which the user might judge the naturalness of a chatbot. These may to some extent be inter-correlated so that a chatbot which does well on one will also do well on one or more of the others.

A Chatbot should be Conscientious. It should be able to keep track of the conversation, attend to the flow of the conversation, maintain themes, pick up appropriate cues, and ask and answer pertinent questions.

A Chatbot should Display Originality. It should have some interesting information about the conversational theme (*specialized* in addition to *common* topics), and it should be able to take the initiative in conversations at times, perhaps by suggestions to change to related themes.

A Chatbot should Display Manners. It should show that it has good conversational habits, can do damage control if a conversation seems to be losing its way, and should maintain an appropriate (perhaps friendly) personality and develop a relationship with the user.

A Chatbot should be Thorough. It should use appropriate grammar and spelling consistently, and consistently adopt an appropriate linguistic register with the user.

As to how such a perfect chatbot should be coded, we are quite agnostic on this point and proponents of the three major approaches to design (Strong Physical Symbols System, Connectionist, or Situational Semantics) will have their own solutions. We favour the Situational Semantics/ Subsumption architecture of Sam and note that the amount to which each of the dimensions is incorporated in each layer will vary with respect to what that layer does; but that overall, the machine should incorporate all four dimensions. This, after all, is what our users tell us they expect.

References

1. Aron, J.: How innovative is Apple's new voice assistant, Siri? *New Scientist* 212 (2386), 24 (2011)
2. Barwise, J., Perry, J.: *Situations and Attitudes*. Mass: MIT Press, Cambridge (1983)
3. Binet, Etude Expérimentale de l'Intelligence. Schleicher Frères & Cie, Paris (1903)
4. Brooks, R.: Intelligence without representation. *Artificial Intelligence* 47, 139–159 (1991)
5. Charette, R.: WellPoint Hires IBM's "Dr." Watson (April 11, 2013), <http://spectrum.ieee.org/riskfactor/biomedical/diagnostics/wellpoint-hires-ibms-dr-watson>
6. Charmaz, K.: *Constructing Grounded Theory*. Sage, London (2006)
7. Cohen, P.: If Not Turing's Test, Then What? *AI Magazine* 26(4), 61–67 (2005)
8. Flanagan, J.: The critical incident technique. *Psychological Bulletin* 51(2), 327–358 (1954)
9. Fodor, J.: *The Language of Thought*. Thomas Crowell, London (1975)
10. French, R.M.: Subcognition and the limits of the Turing Test. *Mind* 99, 53–65 (1990)
11. Hodges, A.: *Alan Turing: the Enigma*. Burnett, London (1983)
12. Hung, V., Elvir, M., Gonzalez, A.J., DeMara, R.F.: A Method For Evaluating Naturalness in Conversational Dialog Systems. In: *Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics*, pp. 2647–2652 (2009)
13. ISO/IEC, *Ergonomic Requirements for Office Work with Visual Display Terminals (VDT)s - Part 11 Guidance on Usability*. ISO-IEC, Geneva (1998)
14. Kirakowski, J., O'Donnell, P., Yiu, A.: Establishing the hallmarks of a convincing chatbot-human dialogue. In: Maurtua, I. (ed.) *Human- Computer Interaction, In-Teh., Vukovar (2009)*

15. Rzepka, R., Ge, Y., Araki, K.: Naturalness of an utterance based on the automatically retrieved commonsense. In: Proceedings of the Nineteenth IJCAI (2005)
16. Semeraro, G., Andersen, H.H.K., Andersen, V., Lops, P., Abbattista, F.: Evaluation and validation of a conversational agent embodied in a bookstore. In: Carbonell, N., Stephanidis, C. (eds.) UI4ALL 2002. LNCS, vol. 2615, pp. 360–371. Springer, Heidelberg (2003)
17. Shawar, B.A., Atwell, E.: Chatbots: Are they Really Useful? LDV Forum 22 (1), 29–49 (2007)
18. Shieber, S.M.: Does the Turing Test Demonstrate Intelligence or Not? Proceedings of the 21st National Conference on Artificial Intelligence 2, 1539–1542 (2006)
19. Turing, A.: Computing machinery and intelligence. *Mind* 59, 433–460 (1950)
20. Weizenbaum, J.: ELIZA – A Computer Program for the Study of Natural Language Communication Between Man And Machine. *Communications of the ACM* 9(1), 36–45 (1966)

Grounding and Turn-Taking in Multimodal Multiparty Conversation

David Novick and Iván Gris

Department of Computer Science, The University of Texas at El Paso, USA
novick@utep.edu, ivangris4@gmail.com

Abstract. This study explores the empirical basis for multimodal conversation control acts. Applying conversation analysis as an exploratory approach, we attempt to illuminate the control functions of paralinguistic behaviors in managing multiparty conversation. We contrast our multiparty analysis with an earlier dyadic analysis and, to the extent permitted by our small samples of the corpus, contrast (a) conversations where the conversants did or did not have an artifact, and (b) conversations in English among Americans with conversations in Spanish among Mexicans. Our analysis suggests that speakers tend not to use gaze shifts to cue nodding for grounding and that the presence of an artifact reduced listeners' gaze at the speaker. These observations remained relatively consistent across the two languages.

Keywords: Dialog, proxemics, gaze, turn-taking, multicultural, multiparty.

1 Introduction

Most studies of multimodal grounding and turn-taking have been based on analysis of dyadic conversation (e.g., [1, 2]). Others have tackled grounding and turn-taking in multiparty conversation, but this was typically either not multimodal (e.g., [3]) or was approached from a theoretical rather than an empirical perspective (e.g., [4]). In the present study, we apply a conversation-analytic approach to begin understanding the mechanisms of grounding and turn-taking in multimodal multiparty conversation. In this study, our principal objective was to explore the empirical basis for multimodal conversation control acts in multiparty conversation, such as those discussed in ([4]). We were interested in questions such as:

- Do grounding behaviors such as nodding get cued in ways similar to those observed (e.g., by [1] and [5]) in dyadic conversation?
- How do the mechanisms of turn-transitions function?
- Does the presence of an artifact lead to changes in grounding behaviors?
- How, if at all, do these behaviors differ across cultures?

We contrasted our multiparty analysis with an earlier dyadic analysis [5] and, to the extent permitted by our small samples of the corpus, contrasted (a) conversations where the conversants had an artifact (a plush toy that they were tasked with naming)

and or did not have an artifact, and (b) conversations in English among Americans with conversations in Spanish among Mexicans.

2 Background

Research on multimodal multiparty conversation is conducted from multiple perspectives, including observing interaction, describing the conversational functions necessary for effective interaction by conversational agents, and implementing these behaviors in agents. The observational perspective, through a discourse-analytic approach, can provide a systematic account of grounding and paralinguistic behaviors in conversation. For example, gaze patterns in multiparty conversation have been analyzed statistically, providing a detailed account of the frequency of different gaze patterns associated with turn-taking [6]. This has led to a probabilistic model of interaction that has been validated empirically, but the model's realism might be considered validated at a descriptive level rather than at a causal level. Such a probabilistic model could lead to relatively reliable functioning of, for example, an automatic gaze-dependent video editor for recordings of conversations [7]. Observational studies of multiparty conversation have also described the role of gesture, beyond gaze, in the process of interaction. For example, conversants' gestures, both head and hand, appear to be a function of the conversant's conversational role and the dialog state; conversants clearly coordinate their utterances and gestures, and this may relate to task structure [8]. However valuable, models produced by discourse-analytic studies do not necessarily provide a deep explanation of how the gaze and turn-taking functions actually work. Going beyond surface simulation—even if highly plausible and effective—requires understanding the specific functional mechanisms for, and the context-specific purposes associated with, conversants' use of paralinguistic behaviors.

While the mechanisms of human-human multiparty conversation management may remain only partially understood, the need for them is clear. The kinds of conversational roles and the broad functions of conversational management needed for effective interaction by embodied conversational agents have been comprehensively catalogued [4]. The functions of interaction management include turn management, channel management, thread/conversation management, initiative management, and attention management. Models of some paralinguistic behaviors have been validated through implementation in conversational agents. For example, a multiparty gaze model based on the findings of Argyle and Cook [9] was validated through simulation in an embodied conversational agent [10].

Whether modeled based on observation or validated through simulation, some aspects of paralinguistic behaviors and dialog management in multiparty conversation have conversational functions that are relatively clear. Other aspects remain confirmed but unexplained from the standpoint of conversational function. For example, conversants in multiparty interaction use a great deal of overlap of utterances [11], but the functional reasons for the overlap are not yet clear. Indeed, a multiparty conversation may actually involve multiple simultaneous conversations, and the conversants must accordingly manage multiple simultaneous conversational

floors [12]. The most plausible account of this management involves different types of conversational moves of splitting the conversation (“schisiming”) or bringing separated threads back together (“affiliating”). These moves can be categorized as schism-inducing turns, schisiming by aside, affiliating by turn-taking, or affiliating by coordinated action [12].

Finally, we note that discourse-analytic comparison of multiparty and dyadic conversation has indicated that differences as a function of the number of conversants vary across cultures [13]. Thus while the use of gaze to coordinate turn-transition differs between speakers of American English and of Mexican Spanish, for Americans this process is a function of group size: gaze plays a relatively smaller role in Mexican multiparty conversation than it does in American [13]. To explore the specific functional mechanisms for conversants’ use of paralinguistic behaviors in multiparty conversation, in this study we oriented our study around four principle issues: whether grounding behaviors such as nodding get cued in ways similar to those observed in dyadic conversation, how the mechanisms of turn-transitions actually function, whether the presence of an artifact leads to changes in grounding behaviors, and how, if at all, these behaviors differ between speakers of American English and of Mexican Spanish.

3 Methodology

To address these questions, we conducted conversation analyses of four 20-second excerpts of conversations from the UTEP-ICT Cross-Cultural Multiparty Multimodal Dialog Corpus [14]. The corpus comprises approximately 20 hours of audiovisual multiparty interactions among dyads and quads of native speakers of Arabic, American English and Mexican Spanish. The subjects were recruited from local churches, restaurants, on campus, and through networks of known members of each cultural group in the El Paso area, which borders Mexico and has, in part because of the university, many representatives of other nations and cultures. In the present research, we focused on interaction in quads of Spanish and English speakers. And because we were particularly interested in grounding and turn-taking, we based our analysis on conversations that had multiple turns over a short period of time.

Tasks 1, 4, and 5 were mainly narrative tasks, where the participants can take turns relating stories or reacting to the narratives of others. Tasks 2 and 3 were constructive tasks, in which the participants must pool their knowledge and work together to reach a group consensus. Tasks 3 and 4 were designed to have a toy provide a possible gaze focus other than the subjects themselves, so that gaze patterns with a copresent referent could be contrasted with gaze patterns without this referent. Task 5 was meant to elicit subjective experiences of intercultural interaction. For each of the four excerpts that formed the basis of the study reported here, we transcribed the speech and annotated the gaze, nods and upper-body gestures of the four conversants in the conversation. Timings were noted with the Elan Linguistic Annotator [15]. From the observed behaviors we then attempted to produce a plausible explanation of how these actions served the conversants in grounding (or not) each other’s contributions to the conversation and in taking conversational turns.



Fig. 1. Speakers of Mexican Spanish conversing, with artifact



Fig. 2. Speakers of American English, conversing, without artifact

4 Results

We begin with our analysis of the conversations conducted by the speakers of American English. In the conversation without the artifact, we observed that the relationship between gaze and nod differed markedly from that observed [5] in dyadic conversation. In dyadic conversation, the listener's nods are typically cued by gaze from the speaker. From a functional standpoint, this represents the listener's providing grounding feedback at points where the speaker can perceive it—and even if the speaker can see the listener peripherally the speaker in effect gives the listener the opportunity to ground (or not to ground) where, presumably, the speaker wants to check on the listener's understanding. Consequently, for speakers of American English, conversants in dyadic conversations looked at each other less frequently than did conversants in multiparty conversations, presumably because higher rates of mutual gaze would be providing too-frequent cues for grounding feedback.

4.1 American Non-artifact Conversation

The (non-artifact) multiparty excerpt we studied here (Table 1 shows the first 12 seconds of the transcript) contains some paralinguistic behaviors associated with dyadic conversation. In particular, most of the listeners are looking at the speaker. This is true as the excerpt starts, where Conversant B is talking, and Conversants A, C and D are looking at B. And it is true again as C takes the floor (at 04:15), and A, then B, and then D look at C. Additionally, B looks away while talking. But other behaviors differ markedly from those expected in dyadic conversation. When B ends the turn it is because C has grabbed the floor; B has not shifted gaze to a listener to check for grounding or to offer the turn. When C begins talking, she looks at A, who does not provide feedback, and then looks away. Although A does engage in a series of small nods while B is talking—but not looking at A, neither A nor the other conversants provide further grounding feedback in this excerpt. Relative to the dyadic conversants observed in [5], these four conversants do very little nodding. However, the conversants do use gestures to reinforce their verbal production. At 00:00 B mimics using a cell phone, and at 08:00 C gestures to indicate a lane change.

Table 1. Excerpt of transcript of American English conversation, without artifact

Time Start	Time End	A Verbal	A Non-verbal	B Verbal	B Non-verbal	C Verbal	C Non-verbal	D Verbal	D Nonverbal
Initial			arms behind back, looking at B		arms crossed, looking away		hands in back pockets, looking at B		R arm crossed, L hand on chin, looking at B
00:00	04:10			when you're driving and you see people talking on the phone and driving	gestures phone with left hand and crosses arms again				
00:00	02:00		succession of small nods						
04:15	05:10					but they're driving like stupid			
04:25	09:50		looks at C						
05:50	07:50				looks at C	they're driving very slow and like	looks at A		
06:00	10:30								looks at C
08:00	10:75					they won't change lanes right	looks away, mimics changing lane		
08:15	08:40				glances at A, looks back at C				
09:40	11:90			-- go off or something or...					

4.2 American Artifact Conversation

In the conversation among speakers of American English with a task that involved the plush-toy artifact (see Table 2), conversants tended to gaze more at the artifact than at each other. At the start of this excerpt, part-way into the conversation, the conversants are all looking at the artifact. And when Conversants B and C shift their gaze away from the artifact, at 04:15 and 06:50 respectively, they look primarily at non-speakers. None of the conversants nods. At 02:10 the turn transition between B and C has a brief overlap, and the transition is not coordinated with a gaze shift. Rather, the gaze shift lags the turn change. Relative to the non-artifact conversation, the conversants use far fewer gestures, perhaps because the conversants' common focus on the artifact takes their attention away from their conversational partners' possible gestural displays.

Table 2. Excerpt of transcript of American English conversation, with artifact

Time Start	Time End	A Verbal	A Nonverbal	B Verbal	B Nonverbal	C Verbal	C Nonverbal	D Verbal	D Nonverbal
Initial			looking at toy on floor; arms crossed		looking at toy on floor; arms crossed		looking at toy on floor; arms crossed		looking at toy (on the floor); right arm crossed; left arm on chin
00:00	00:24					[laugh]			
00:24	02:17			the anonymous beast					
02:10	05:30							it has to be named or else how'll people tell other people what to buy	
04:15	06:20				looks at A, then C				
06:50	08:40						looks at A		

4.3 Mexican Artifact Conversation

We now turn from the excerpts of the American conversations to the excerpts of the Mexican conversations. In their conversation that included the plush-toy artifact (see Table 3), Mexican participants generally nodded when the speaker gaze was focused on the artifact. While nodding was below the overall frequency compared to conversations when no artifact was involved, when listeners did nod it was after a verbal consensus and agreement and rarely otherwise. For example, in the artifact task after speaker B proposes a toy name to the listeners at 16:50, A immediately takes the turn within a half a second and verbally agrees three times in succession with 1.5-2.0 second intervals. Meanwhile, C produces a succession of small nods between each verbal statement, even though the four conversants have their gaze focused on the artifact. This seems similar to the nodding behavior of Conversant A in the American non-artifact conversation. But these behaviors appear at odds with the explanation in [5], which suggested that if the speaker is not looking at you, it does not do much for you to nod because the speaker may not (cf. peripherally) see your action. It is important to note, though, that there were no artifacts involved in that study. One possible explanation for nodding, even when no one is looking, can be the need to express agreement while not wanting to take the floor to express it. The task asked for a group consensus on the naming of the artifact, which required all participants to agree on a name. Silence may be a weak agreement that is reinforced by non-verbal behavior to help achieve the task.

We also note turn-taking differences for Mexican conversants in conversations with and without an artifact. When an artifact was involved, all the conversants looked at the plush toy through the majority of the conversation, looking away and at other participants only once (each) in the 20-second transcript. This occurred during seconds 15-17, when (B) suggested a toy name, but none retained the gaze more than

a second. While not producing non-verbal behaviors, participants instead signaled turn taking by repeating a previous statement in as-a-matter-of-fact intonation or by adding to the collective description of the artifact with simple sentences (“es azul / it’s blue”, “tiene colmillos / it has fangs”, “tiene cuernos / it has horns”). In general, repetition seems to act as an acknowledgment and an invitation for someone else to take the floor, as the repeater rarely adds anything else to the conversation. Indeed, this seems like the “display” method of grounding described by Clark and Schaefer as the strongest form of acceptance of a contribution to discourse [16].

Table 3. Excerpt of transcript of Mexican Spanish conversation with artifact

Time Start	Time End	A Verbal	A Nonverbal	B Verbal	B Nonverbal	C Verbal	C Nonverbal	D Verbal	D Nonverbal
15:00	16:50		(looking at toy) small step backwards	blue punk no, algo asi?	(looking at toy) touches toy's hair		(looking at toy) quick glance at B		(holding toy, looking at toy)
16:50	17:00	blue punk, si	quick glance at B				looks at toy		
17:00	17:50			blue punk	quick glance at A				
17:50	18:00			es azul					
18:00	19:00			y luego trae el pelo aca					
19:00	19:50	si blue punk		y luego			nods		
19:50	20:00			son los punk					
20:00	20:50			blue punk			nods		
21:00	22:00	si blue punk							
22:00	23:00						nods	ey, blue punk	
23:00	24:00	blue punk							

4.4 Mexican Non-artifact Conversation

In the Mexican non-artifact section of the corpus (see Table 4), conversants were relatively more inclined to gesture. This reinforces the effects of gaze, in which if you are being observed motivates the speaker to enhance his/her conversation with gestures. These gestures can be separated in two different types. The first type is analogous to the ones found in the artifact conversations, which are used for agreement or acknowledgement. One of the main differences is that without an artifact, the gesture tends to be done in conjunction to the verbal statement. This can be observed from 02:00 – 08:00 on (B), (C) and (D) in different statements, usually briefly following or in conjunction with “si” or “a mi tambien” (yes / me too). The second gesture type is exclusive to the non-artifact participants. In this case, conversants use gestures to enact the verbal part. In this particular conversation the conversants are critiquing the movie Titanic. Hand gestures conversants use include imitating the sinking ship (A) at 07:30, a necklace (D) at 12:00, and people drowning (B) at 15:00.

Overall, gestures appear to be significantly more frequent when there is no artifact. Without the artifact, conversants appeared to prefer overlapping for turn taking. However, conversants rarely used overlap to change topic but rather to elaborate on the current topic. During the first thirteen seconds of the conversation, verbal agreement was used to change turns, although no one kept the floor for long, and the ideas proposed after taking the floor were left unresolved. For example, at 12:00 conversant D takes the turn for the first time within seven seconds with a contribution but expresses only an unfinished sentence "... y luego se quitan el collar y ..." (and then they take the necklace off and...).

Table 4. Excerpt of transcript of Mexican Spanish conversation without artifact

Time Start	Time End	A Verbal	A Non-verbal	B Verbal	B Non-verbal	C Verbal	C Non-verbal	D Verbal	D Non-verbal
06:50	08:00				looks at D				
07:30	08:10	si y luego como el barco	hand gestures indicating the sinking ship				locks hands at waist to her front		looks away from group
08:00	11:50				looks at A				
08:20	08:80	como osea							
08:30	14:00								looks at A
09:00	09:30	Si							
09:00	12:50		looks at B						
10:00	11:00	como va cambiando y todo eso							
10:80	11:20			si	nods				
12:00	13:00							y luego que se quitan el collar y...	moves in her place and gestures a necklace with both hands
12:50	14:00		looks at D						

5 Conclusion

Based on our observation of the four conversational excerpts discussed in Section 4, we now return to the four questions that motivated our study.

Do grounding behaviors such as nodding get cued in ways similar to those in dyadic conversation? The evidence in the multiparty conversations we studied suggests that multiparty conversants nod less frequently and even these fewer nods are not being cued by the speaker’s gaze shift.

How do the mechanisms of turn-transitions function? The evidence, which includes greater overlap at turn boundaries, suggests that conversants in multiparty conversation do not rely as much on gaze as a turn cue as do conversants in dyadic

conversation. Rather, multiparty conversants repeatedly overlapped at turn boundaries, especially where one party grabbed the floor—possibly because the conversant could not engage the speaker’s gaze.

Does the presence of an artifact lead to changes in grounding behaviors? The evidence suggests that the presence of an artifact draws the conversants’ gaze, thus reducing the amount of time that listeners gaze at the speaker. This may contribute to the phenomenon of speakers tending not to use gaze shifts to cue nodding as a grounding behavior. In one case (Mexican, artifact), the conversants seemed to substitute display for continued attention as a grounding behavior.

How, if at all, do these behaviors differ across cultures? While we found some differences between the behaviors of speakers of American English and of Mexican Spanish, these differences likely reflect natural variation in conversation rather than clear cultural differences. Rather, comparison of the conversations across cultures revealed similarities, particularly with respect to differences in gaze patterns across the artifact/non-artifact condition and with respect to the lack of cueing of nods.

References

1. Nakano, Y.I., Reinstein, G., Stocky, T., Cassell, J.: Towards a model of face-to-face grounding. In: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, vol. 1, pp. 553–561 (2003)
2. Bavelas, J.B., Coates, L., Johnson, T.: Listener Responses as a Collaborative Process: The Role of Gaze. *Journal of Communication* 52, 566–580 (2002)
3. Branigan, H.: Perspectives on Multi-party Dialogue, *Research on Language and Computation*, vol. 4, pp. 153–177 (2006)
4. Traum, D.: Issues in multiparty dialogues, *Advances in agent communication*, pp. 201–211. Springer (2004)
5. Novick, D.: Paralinguistic behaviors in dialog as a continuous process. In: Proceedings of the Interdisciplinary Workshop on Feedback Behaviors in Dialog, Stevenson, Washington, September 7-8, pp. 54–57 (2012)
6. Otsuka, K., Takemae, Y., Yamato, J.: A probabilistic inference of multiparty-conversation structure based on Markov-switching models of gaze patterns, head directions, and utterances. In: Proceedings of the 7th International Conference on Multimodal Interfaces, pp. 191–198 (October 2005)
7. Takemae, Y., Otsuka, K., Mukawa, N.: Impact of video editing based on participants’ gaze in multiparty conversation. In: CHI 2004 Extended Abstracts on Human Factors in Computing Systems, pp. 1333–1336 (2004)
8. Battersby, S.A.: Moving together: The organisation of non-verbal cues during multiparty conversation. Doctoral dissertation. University of London, Queen Mary (2011)
9. Argyle, M., Cook, M.: *Gaze and Mutual Gaze*. Cambridge University Press, London (1976)
10. Gu, E., Badler, N.: Visual attention and eye gaze during multiparty conversations with distractions, *Intelligent Virtual Agents*, pp. 193–204. Springer (2006)
11. Shriberg, E., Stolcke, A., Baron, D.: Observations on overlap: Findings and implications for automatic processing of multi-party conversation. *Proc. Eurospeech* 2, 1359–1362 (2001)

12. Aoki, P.M., Szymanski, M.H., Plurkowski, L., Thornton, J.D., Woodruff, A., Yi, W.: Where's the party in multi-party?: Analyzing the structure of small-group sociable talk. In: Proceedings of the 2006 20th Anniversary Conference on Computer-Supported Cooperative Work, pp. 393–402 (2006)
13. Herrera, D., Novick, D., Jan, D., Traum, D.: Dialog behaviors across culture and group size. In: Stephanidis, C. (ed.) Universal Access in HCI, Part II, HCII 2011. LNCS, vol. 6766, pp. 450–459. Springer, Heidelberg (2011)
14. Herrera, D., Novick, D., Jan, D., Traum, D.: The UTEP-ICT Cross-Cultural Multiparty Multimodal Dialog Corpus. In: Multimodal Corpora Workshop: Advances in Capturing, Coding and Analyzing Multimodality (MMC 2010), Valletta, Malta (2010)
15. Tacchetti, M.: ELAN User Guide, version 4.1.0 (2011), http://www.mpi.nl/corpus/html/elan_ug/index.html
16. Clark, H.H., Schaefer, E.F.: Contributing to discourse. *Cognitive science* 13(2), 259–294 (1989)

Situated Multiparty Interaction between Humans and Agents

Aasish Pappu, Ming Sun, Seshadri Sridharan, and Alex Rudnicky

Carnegie Mellon University, Pittsburgh PA 15213, USA

Abstract. A social agent such as a receptionist or an escort robot encounters challenges when communicating with people in open areas. The agent must know not to react to distracting acoustic and visual events and it needs to appropriately handle situations that include multiple humans, being able to focus on active interlocutors and appropriately shift attention based on the context. We describe a multiparty interaction agent that helps multiple users arrange a common activity. From the user study we conducted, we found that the agent can discriminate between active and inactive interlocutors well by using the skeletal and azimuth information. Participants found the addressee much clearer when an animated talking head was used.

1 Introduction

When more than two people engage in a human-human conversation, they seamlessly communicate and sense who is speaking, who desires to speak, and who is simply listening. This phenomenon is referred to as conversation management in a multiparty scenario. Both verbal and non-verbal cues enable efficient conversation management. In a typical multiparty conversation, participants share a common floor of speech and take turns one at a time to address the floor or a particular addressee. Humans not only show efficient floor management skills but also conform to norms such as politeness and social-role in such situations. On the other hand, imagine a situation with one of the participants being an artificial agent (robot). Here, the problem of floor/conversation management escalates. [1] presented issues that arise in such a situation. They discussed the issues related to a) Participant roles b) Interaction management and c) Grounding and Obligations. Based on these issues, we address three research questions 1) How does an agent determine the roles of different participants? 2) When is appropriate for the agent to take/release the floor? 3) How does an agent communicate (and ground) its understanding of floor and conversation dynamics?

The dynamics of a multiparty conversation are distinct from a two-way conversation (or dialog). [2] proposed special conversation acts called hearer and speech acts. They argue that the traditional definition of a speech act only explains the act of a speaker addressing (assert, promise or apologize) an addressee. Here the assumption is that all addressees are hearers. But, in a multiparty conversation all hearers i.e., all listening participants are not necessarily addressees.

From a computational perspective, we can define conversation as an act that encompasses multiple dialogs with shared or distinct goals with interlocutors who share the floor by taking turns. In a conversation, an agent either jointly informs or requests while addressing all the participants. Whereas in a dialog, an agent only communicates with a particular addressee while all other participants assume the hearer role.

Detecting active participants in a conversation is challenging because one can join or leave the conversation anytime. An agent should keep track of who might engage in a conversation and who has disengaged from a conversation. In addition, it should maintain topic congruity with the active participants on the floor. Therefore, it should decide whether to include or exclude a non-participant in the middle of an active conversation. [3] found that rule-based addressee detection methods are comparable to that of supervised statistical methods such as bayesian networks on a mulitmodal meeting corpus. Gaze patterns, speech, and gesture [4] [5] were found to have predictive power to build addressee detection models. Combining different multimodal inputs compensate for the drawbacks of individual modalities. In our work, we use skeleton and auditory information generated by a Microsoft Kinect[6] to tackle the attention detection problem.

Once the conversation is active, the agent needs to constantly monitor who has the floor and who might get the floor. If the agent makes a request for the floor, it needs to know and monitor how many participants may take the floor before it get its turn. [7] have proposed a heuristic turn-taking policy in a multiparty scenario in a social setting. Such a policy sets up a default behavior for the system and is helpful until the system acquires sufficient data to learn a decision model through interactions.

Since conversation dynamics are complex, an agent should effectively convey its understanding of the dynamics to everyone else in the conversation. In a dialog, an agent only needs to communicate whether or not it understood the user's utterance. In a multiparty conversation, it should also communicate its own understanding of the floor ownership. This helps the participants to take their turn and open up the floor in a timely fashion. An animated talking head or a face has been a norm for embodied agents [8]. Both robotic heads and projection on 2D flat panels have been used as a solution to non-verbal communication, although [9] argue that in multiparty scenario, projections on 2D panels are insufficient to convey which user is being addressed. Instead, they propose a 3D animated back-projected avatar with a mechanical tilt-able neck. In this work, we use animated line-drawings as a 2D talking head. We conducted user studies to investigate its efficacy in turn-taking and state transparency.

In this work, we describe a new framework for multiparty conversation management for an agent in a social settings. This framework tries to address the three fundamental challenges described above. The agent detects the active participants with the help of skeletal and audio sensory information and engages them in conversation. Then, it uses this sensory information and the current conversation state to actively monitor which participant has the floor. With the help of verbal and non-verbal cues (gaze), it conveys its belief of floor

ownership and utterance understanding. In addition to addressing these challenges, we also made this framework extend to existing dialog applications for multiparty applications.

This paper is organized as following: In Section 2, we discuss the architecture of the conversation framework. In Section 3, we present empirical results evaluating a system built on this framework. Finally, in section 4 we present concluding remarks and future directions of this work.

2 System Architecture

2.1 Ravenclaw/Olympus Framework

The agent (“SocBot”) is implemented using the Olympus/Ravenclaw dialog framework, augmented with multi-modal capabilities (see gray blocks in Fig. 1). A Kinect device is used to acquire speech and human skeleton data. Skeleton information and sound source azimuth information are used to manage the agents attention strategy and as part of the voice activity detection (VAD) process. Speech is decoded by the Automatic Speech Recognizer (ASR) and then processed by a semantic parser. It is further processed by an Input Confidence Estimator (ICE) that combines language, skeleton and azimuth information to determine a given inputs intentionality. Depending on the user input and the current context, the Dialog Management (DM) component decides the next action; this may involve communicating with the Domain Reasoner (DR). Finally a natural language response is generated by the NLG and systems response is synthesized via text-to-speech engine (TTS). Interaction manager (IM) coordinates between system’s listening and speaking states to allow barge-ins from the user.

2.2 Customized Ravenclaw/Olympus Framework

We adapted the Olympus/Ravenclaw framework[10] to handle multiparty interaction. For each user, a set of essential components (ICE and DM, as shown in gray blocks on the right side in Fig.1) are spawned and interconnected. Three new components are implemented to naturally handle multiparty conversation (see black blocks in Fig.1). Awareness Server tracks the users in front of the agent and associates each speech input with its corresponding user by using both the skeleton and azimuth information. Conversation Manager (CM) decides when to speak, what to speak, when to listen and whom to listen to. Talking head extracts the addressee stream from CM and NLG. Through its behavior, it displays the current system state (e.g., understood, confused) and the current focus of the agent.

2.3 Conversation Manager

Conversation Manager can access information of all on-going dialogs. CM gates the message flows from speech signal to each dialog manager and natural language requests from dialog managers to synthesizer. To control when to speak

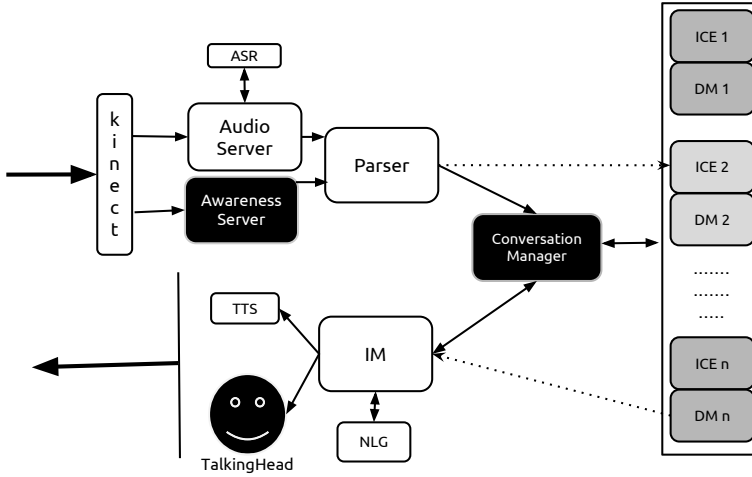


Fig. 1. Multimodal, Multiparty Ravenclaw-Olympus

and what to speak, it also mediates between Dialog Manager (DM) and IM in the output direction. Therefore, CM knows which DM wants to take the floor at any given time. Once the output requests are sent from both Dialog Managers to CM, CM interprets the semantics of the requests and decides which one (user A’s, user B’s, or a combined one) to forward to IM according to the dialog state. To control when to listen and whom to listen to, CM mediates between Interaction Manager (IM) and Input Confidence Estimator (ICE) in the input direction. Only when CM receives parse(s) from expected user(s), it will send the input to the expected ICE(s) which will then pass the message on to the DM(s) accordingly.

For example, in a dialog state where user hobbies are required, both Dialog Managers send NLG requests with different parameters (DM-A: requests A’s hobby. DM-B: requests B’s hobby.) to CM. CM first realizes that the two requests are of the same dialog state. Then it finds whether the parameter (user name) is different for these two requests. If so, it generates a plural version of the request (request A and B’s hobbies). In the input phase, since CM knows that two answers from two users are expected, it will not send the parses of the input to the rest of the system until two utterances from these two users are provided. Note that a timeout can be triggered if CM does not receive two inputs within a period of time.

In human-human multiparty conversation, we use verbal and non-verbal cues to express our focus. For example, if one is expecting a response from person A, one would either say “A, what about you?” or look at person A. instead of someone else. Similarly, these types of cues are provided by CM as well. In the earlier output example, CM interprets the NLG requests and attaches

the addressee (user name(s)) to the system prompt. In our system, CM also propagates the addressee information to the talking head, which looks at the expected user(s) accordingly. Details of the talking head are described below.

2.4 Awareness Component

To hold a multiparty conversation, two issues need to be addressed: 1) when to engage and disengage 2) who is the active speaker right now. In our system, the number of skeletons in front of the agent is tracked and updated all the time. When no humans are about, the agent is idle. The agent wakes up upon seeing skeleton(s) in the environment. Based on the number of skeletons available, it decides whether to start a single party dialog or a multiparty conversation. After establishing a conversation with the users, each speech input is associated with the appropriate user. If a skeleton appears or leaves the environment during the conversation, CM is notified to trigger certain conversation-level actions such as acknowledging a newcomer or suspending the dialog channel with the user that left.

To fulfill the capabilities described above, we use the information from the Kinect sensor's microphone array and the visual sensors on-board. This set-up allows multiple users to engage in hands-free fluid interaction with the agent, without wearing close-talking microphones.

As a fundamental requirement to have conversation with multiple people, the system needs to perceive three different types of events. Firstly, it needs to capture audio from mobile sound-sources a few feet away. For this purpose, the far-field microphone array in the Kinect acts as the audio sensor. Before streaming the audio packets to the VAD and the recognizer, we enable on-board noise suppression and echo cancellation to obtain a clean signal. The audio gain level is dynamically adjusted to suit the environmental changes and volume-levels of independent users.

Secondly, the system needs to be aware of user-engagement events. It needs to detect when a user joins the floor, leaves the floor, and should avoid reacting to nonparticipants. We use the skeleton tracking capability of the Kinect to scan the environment for skeletons every 30 frames a second. This allows us to detect skeletal events and assign a unique person-id to each skeleton by tracking/monitoring the skeleton in the environment. To avoid premature firing of events in cases of passerby skeletons entering the environment or the sensor dropping skeletons momentarily, the Awareness component waits for a few hundred milliseconds to observe the environment before making the decision to fire a particular event.

Thirdly, the system needs to be able to discern whom a particular speech input came from. For this purpose, we use the microphone array to track the audio beam, monitoring for changes in the angle of the sound source. When voice is detected at the audio-signal level, we look for a matching skeleton for the current sound-azimuth. When the difference between the orientation of the closest skeleton and the azimuth is within a 15-degree angle, we tag the decoded

result for that particular utterance with the user-id of the matched skeleton. The tagged input is sent to the CM for appropriate action.

Since the system knows the exact association between the skeletons and the Dialog Managers, it knows which direction to look at for a particular user's input. This knowledge of the user orientation is used by the graphical user interface (talking head) to direct the prompt towards the target user(s). For example, when DM-A wants an input, the agent looks towards user A's direction, prompt user A, and later expects an utterance from user A.

2.5 Talking Head

As an important non-verbal cue to explicitly indicate the current focus of the agent, we implemented an animated talking head that gazes at whoever it is expecting a response from. Addressee information is sent from CM and the orientation of each user is provided by awareness component. The talking head would also move its lips when the agent is speaking. The following scenarios are considered.

- When addressing the floor and none of the users has replied, its eyes scan the two users.
- When addressing the floor and one of the users has already replied, it looks at the user who has not replied yet.
- When addressing one user, it looks at that user.
- Upon receiving an input not from either of the users, it looks confused.

3 Social Behavior of Multiparty Conversation

The goal of the agent is to engage multiple users in a conversation. To be natural and social, we designed the agent to handle the following situations which are likely to occur in a daily situations.

3.1 Multiparty Conversation Scenario

When talking to two persons, the agent needs to make it clear whom it is addressing and from whom it is expecting a respond. In some dialog states, the agent is addressing the floor by verbally using a plural form of prompt (e.g., "What are your names?") and visually moving its eyes towards both users back and forth. Upon receiving a speech input from one of the users, it gazes at the one other user. In some states, the agent will completely focus on one particular user. Again, the agent will keep eye contact with that specific user during those states. That user's name will appear in the prompts as well if available. One example of this conversation is shown in Fig. 3.3.

3.2 Single User Conversation with Intervention

This scenario describes the situation that one user is talking to the agent while someone else shows an intention to converse with the agent as well. In this case, the agent will shift its focus to the new comer and acknowledge that user. After that, the agent comes back to the first user without losing the context. Once the dialog with the first user is finished, a new interaction can be established if the new comer is still willing to converse.

3.3 Multiparty Scenario with Disengagement

When two users are having a conversation with the agent, if any of them leaves, the conversation should still move on. Awareness component knows exactly which user has left. CM suspends the dialog channel of that user. The agent will focus on the remaining user thereafter. In our future work, we want the agent to reopen the suspended dialog if that particular user comes back in a short period of time. How to help that user recall the earlier conversation context is an interesting research question.

<p>Multiparty Conversation Scenario <i>Two users walk to the system</i> <i>System detects the users</i> S: Are you guys together? U1: YES S: What are your names? U1: DOE U2: SMITH <i>System misses U2's utterance</i> <i>turns towards U2</i> S: What is your name? U2: SMITH S: Hey DOE and SMITH, what activity do you want to schedule? U1: HIKING U2: CHESS <i>System initiates a subdialog with U1</i> S: Do you want to try "chess" this time? U1: SURE S: Thanks for agreeing. <i>system addresses jointly</i> S: Your activity has been scheduled.</p>	<p>Single-User Scenario with Intervention <i>System in the middle of a dialog</i> U: DOE S: What is the activity do you want to schedule? <i>Someone else tries to participate in dialog</i> <i>System detects them and acknowledges their presence</i> S: Please give me a minute. I will attend you soon. <i>resumes dialog with DOE</i></p>
<p>Multiparty with Disengagement <i>System in the middle of an interaction</i> U1: DOE <i>U2 leaves before conversation ends</i> <i>System suspends channel with U2</i> S: Hey DOE, what is the activity do you want to schedule? U1: HIKING S: Your activity has been scheduled.</p>	

Fig. 2. Example Conversations

4 Experiment

A user study is conducted to answer the following questions:

- How well is the agent distinguishing multiple users’ input via skeleton and azimuth features?
- Whether the expected user(s) responded to any given question or not?
- Subjectively, is it clear to the users who is the addressee of the agent in any given dialog state?

The user study included 12 multiparty conversations. Each conversation was carried out by two subjects and the agent. Initially, the two subjects stood outside the view of the agent (Red positions in Fig.3 and Fig.4). Then they walked up to the green positions and faced the agent. Two green positions are one meter in front of the Kinect sensor and the talking head, with 0.8 meter between them. Out of the 12 conversations, 6 were with a talking head. In total, 6 subjects are involved. Each subject participated in 4 conversations — 2 with talking head and 2 without.

The goal of each conversation is to let the agent arrange a common activity for the two subjects. The agent would ask their names first and the two subjects were expected to say their names one by one. Then the agent would ask what activity they wanted to pursue. Again the subjects would tell their activities one after another. The order of response did not matter. After knowing names and activities, the agent would convince one of the subjects to try out the activity of the other subject. A final confirmation would be spoken by the agent.

The subjects don’t know in advance who is/are expected to speak in any dialog state. They have to interpret the addressee of current dialog state on their own by language cues (e.g., the agent may say “Hey A, what do you want to schedule for tonight?”), talking head cues (e.g., the talking head will look at the current addressee) and conversation context. The two subjects decide whether to take the floor or not. After each conversation, they filled out a survey form.



Fig. 3. Real view of the setup

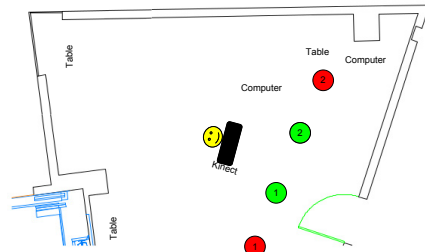


Fig. 4. Schematic view of the setup

4.1 Experiment Results

We found that out of the 140 user utterances, the agent is able to associate 81% of the speech inputs with the correct user. We observed that skeleton angles for both users are very stable. However the azimuth angle was not stable. As a result, sometimes the azimuth cannot be aligned with the correct skeleton, which leads to the errors that the agent either mistook user A’s speech as user B’s (8% of the total utterances) or none of theirs (11%).

To investigate whether the expected user(s) responded in any given dialog state, we accumulated the number of user utterances which were spoken by the wrong user and were discarded by the system. We found that when there was no talking head, 22% of the user input utterances are wasted. When there was talking head, only 8% were from the wrong user. However, the difference is not statistically significant ($p = 0.24$). Further investigation and experiments are required to verify whether significant difference can be observed when more subjects are involved.

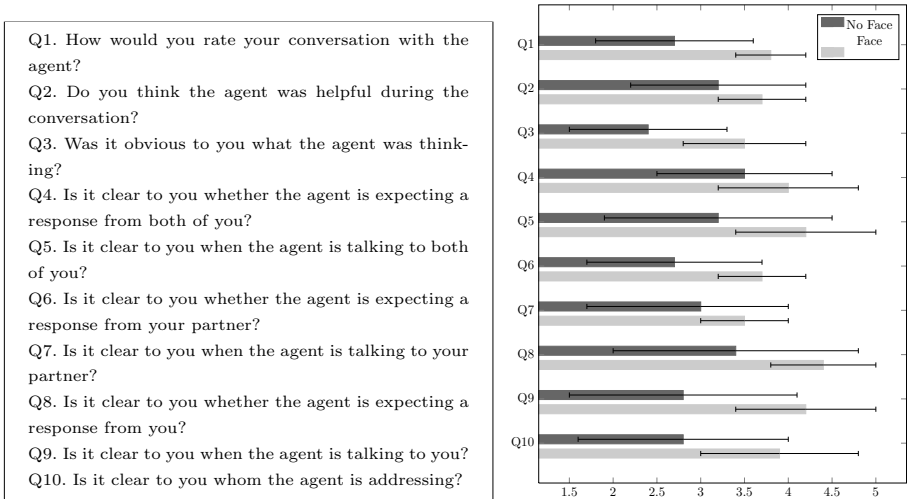


Fig. 5. Subjective results

From the survey results (higher the better), we find out that the overall ratings of the conversation are 2.75 for the agent without a talking head and 3.83 for the talking head one ($p < 0.01$).

The result of the subjective evaluation shows that the addressee is significantly clearer ($p < 0.05$ for each question) when talking head is used. Fig.5 describes the subjective questions and the average score, variance of each question.

5 Conclusion

In this study, we designed and implemented a multiparty spoken dialog system which can simultaneously engage and converse with multiple interlocutors. The

addressee of the agent in any given state is indicated via language cues and talking head facial expression cues. From the results of the user study, we found that the agent is able to discriminate multiple users by using skeleton and azimuth information provided by a Kinect. Subjectively, participants found the talking head agent indicates its focus significantly more clearly than the agent without talking head. These results confirm the intuition that multiple conversational cues support more robust and natural interactions.

References

1. Traum, D.: Issues in multiparty dialogues. *Advances in agent communication, 1954–1954* (2004)
2. Clark, H.H., Carlson, T.B.: Hearers and speech acts. *Language*, 332–373 (1982)
3. Akker, R., Traum, D.: A comparison of addressee detection methods for multiparty conversations (2009)
4. Nakano, Y., Ishii, R.: Estimating user’s engagement from eye-gaze behaviors in human-agent conversations. In: *International Conference on Intelligent user Interfaces*, pp. 139–148. ACM (2010)
5. Bohus, D., Horvitz, E.: Learning to predict engagement with a spoken dialog system in open-world settings. In: *Proceedings of the 2009 Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 244–252 (2009)
6. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from a single depth image. In: *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition* (2011)
7. Foster, M., Gaschler, A., Giuliani, M., Isard, A., Pateraki, M., Petrick, R.: “two people walk into a bar”: Dynamic multi-party social interaction with a robot agent. In: *Proc. of the 14th ACM International Conference on Multimodal Interaction ICMi* (2012)
8. Fukuda, T., Taguri, J., Arai, F., Nakashima, M., Tachibana, D., Hasegawa, Y.: Facial expression of robot face for human-robot mutual communication. In: *Proceedings of 2002 IEEE International Conference on Robotics and Automation*, vol. 1, pp. 46–51. IEEE (2002)
9. Al Moubayed, S., Beskow, J., Skantze, G., Granström, B.: Furhat: A back-projected human-like robot head for multiparty human-machine interaction. In: Esposito, A., Esposito, A.M., Vinciarelli, A., Hoffmann, R., Müller, V.C. (eds.) *COST 2102. LNCS*, vol. 7403, pp. 114–130. Springer, Heidelberg (2012)
10. Bohus, D., Rudnicky, A.: The ravenclaw dialog management framework: architecture and systems. In: *Proceedings of the 2008 Computer Speech and Language*, pp. 332–361 (2008)

Enhancing Human Computer Interaction with Episodic Memory in a Virtual Guide

Felix Rabe and Ipke Wachsmuth

Artificial Intelligence Group, Bielefeld University
Universitätsstraße 25, 33615 Bielefeld, Germany
{frabe, ipke}@techfak.uni-bielefeld.de

Abstract. Have you ever found yourself in front of a computer and asking it aloud: “Why?” We have constructed a cognitively motivated episodic memory system that enables a virtual guide to respond to this question. The guide, a virtual agent based on a belief–desire–intention (BDI) architecture, is employed in a Virtual Reality (VR) scenario where he accompanies a human visitor on a tour through a city. In this paper we explain how the agents memorizes events and episodes according to an event-indexing model and how the interaction is enhanced by using these memories. We argue that due to the cognitively motivated nature of the event-indexing model every interaction situation can be described, memorized, recalled and explained by the agent.

Keywords: Episodic Memory, Event Indexing, Virtual Guide.

1 Introduction

Memories are not just the autobiographical log of yourself, they are the significant foundation that enables humans to plan next actions and let us build an expectation of what might happen in the future. Especially in interaction with other human beings we rely on past episodes with other persons, we adapt to situations based on what we experience and store in episodic memory.

Our work is centered around the virtual humanoid agent *Max*, cf. [5,1]. Max has already lots of skills and is based on a belief–desire–intention (BDI) cognitive architecture. He has proven a helpful interaction partner, e.g. in assisting in construction tasks or “working” as a museum guide, where he is present on a large screen, conducts small talk with visitors and explains sights [7]. But up to now Max has no memory of his own actions and his experiences.

Therefore we have designed a cognitively motivated episodic memory system for Max. We have also conceived a virtual guide scenario, where Max can utilize his new skill in interaction. The episodic memory system is employed when the agent is guiding a visitor through a virtual environment. The main function is memorizing the agent’s actions and interactions with the human visitor. Improved guiding recommendations are generated from similar previous episodes.

This paper is an extended version of [8]. It gives details on the guide scenario and explains how the interactions with an artificial agent are influenced by an episodic memory.

2 Related Work

The concept of episodic memory was first coined by psychologist Endel Tulving in 1972 [11]. Tulving suggested that there are two distinct types of declarative long-term memory: Episodic and semantic memory. While the latter is factual knowledge about the world, episodic memory deals with temporally dated episodes or events and temporal-spatial relations among these events. Every “item in episodic memory” (Tulving) is a more or less faithful record of a person’s experience of an occurrence, and may include the perceptible properties of that moment.

The subject of episodic memory has also attracted interest in computer science, there have been different implementations of computational models of episodic memory. Johnson [4] built an artificial fighter pilot that makes use of episodic memory to explain itself during debriefing. Ho [3] built an autobiographic memory system for an agent to locate resources encountered before and used it in further work to enhance virtual characters, so that they were able to talk about personal past experiences. Tecuci and Porter [10] created a generic memory module for events, where a generic episode has three dimensions: context, contents, and outcome, but only used it for planning. Nuxoll and Laird [6] extended a cognitive architecture (Soar) with episodic memory. They showed that an autonomous agent (a virtual tank) performs better if it can use episodic memory for reasoning. Brom et al. [2] proposed a “full episodic memory” for a non-player character of a computer role-playing game that allows the reconstruction of the character’s personal story.

The foundations for our episodic memory system are the Event-Indexing Model by Zwaan and colleagues [15] and the Event Segmentation Theory by Zacks and Tversky’s [14]. The Event-Indexing Model describes how readers of short stories construct a model of the situation in the text along five indices: **Time**, **Space**, **Causality**, **Intentionality**, and **Protagonists**. These dimensions store answers to the questions of what happened when, where, why and how, and who was involved. We incorporate these five dimensions to describe and index the experiences of the virtual agent. Complementary to this the Event Segmentation Theory defines an event as “a segment of time at a given location that is conceived by an observer to have a beginning and an end” [14]. In further work Zacks et al. [13] found evidence that when human perception does not match the internal prediction, an event boundary is perceived. The Event Segmentation Theory also states that events are organized in partonomic hierarchies, which means they may span very long and very short periods of time, and a long event can embrace several short ones.

3 Events and Episodes

In common language *event* is a broad concept. To narrow it down we define every observable occurrence as an event, in contrast to common language, where especially extraordinary occurrences are called event. Second, we follow the definition

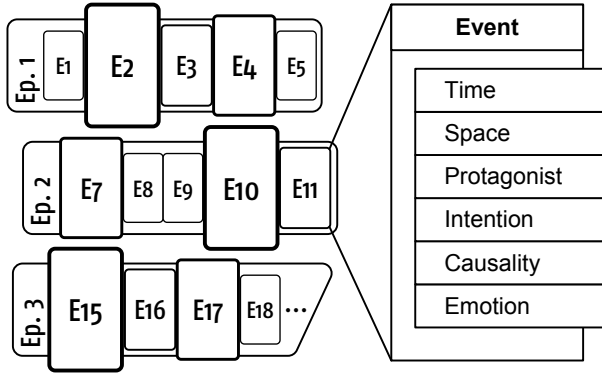


Fig. 1. How episodes and events are conceptualized (as introduced in [8]): Events are grouped into episodes. From the outside of episodes only events with a strong emotional impact are visible, e.g. event E2. The enlarged event E11 shows the six indices. Episode Ep. 3 is the current episode to which new events can be added.

of Zacks and Tversky [14], that an event is a segment of time with a beginning and an end. Third, we are only considering events of equal level, that means we do not consider events that contain other events. In our approach events are not organized in partonomic event hierarchies, but in *episodes*. Fourth, we index events along five dimensions (Time, Space, Protagonists, Intentionality and Causality), according to Zwaan’s event-indexing model. Fifth, we added emotion as sixth dimension, since we may use it as a cue for importance of events. Figure 1 illustrates how episodes group events together and how events are conceptualized.

3.1 Event Indices

We now give a short overview of how the individual indices are conceptualized, for a detailed explanation and the event and episode metrics we created refer to [9].

Time records the time of the event’s begin and end. It also stores the duration of the event which is used for comparison.

Space holds the location of the event in virtual world coordinates. But similar to humans, who normally do not tend to memorize places in GPS coordinates, these are mapped to a named place in the world which is used for comparison.

Protagonists stores named representations of the individuals present during an event. This can be the agent (‘I’) himself and any known and named visitors.

Intention represents the intention the agent has during the current event. Since the agent is based on a BDI cognitive architecture the agent’s next actions are based on its intention.

Causality is what has lead to the event and contains the answer to the question why. This can be either a named percept of the agent (e.g. a request

stated by the visitor to show a certain building), or a named action the agent performed (e.g. completely leading the way to a certain building).

Emotion contains the current emotional state of the agent and may also contain the emotional impulse the agent receives. Events with a high emotional value are considered more significant and are more likely to be remembered.

3.2 Event Boundaries

The segmentation and memorization of events is an “unconscious” mechanism, we follow the idea stated by Zacks et al. [13] that event segmentation is a spontaneous outcome. In particular, whenever Max perceives an input, a change in the environment, or a change of his internal state, the memory system compares this to the current event and if the change is significant enough, a new event is memorized. Max’s perception incorporate dialogue input and other actions of the visitor of the virtual world, e.g. navigation to a location and focusing on a specific virtual item. He also knows where he and the user are, he has knowledge of his intention, his emotional state and the actions he is performing.

4 Guide Scenario

Our scenario comprises a virtual tour guide accompanying and guiding a non-local person through a large and complex virtual environment. A visitor encounters Max in a CAVE-like virtual reality, *Virtual Tübingen*, a virtual city realistically modeled after the historic center of Tübingen in Germany (see Fig. 2).

The virtual model of Tübingen is large and complex, it covers an area of 500x150 m², it has 15 streets which are mostly curved and varying in width, and about 200 houses with different, photorealistic textures. Originally the model was developed at the Max Planck Institute for Biological Cybernetics as a naturalistic, controllable environment for investigating human spatial cognition [12]. We extended it by tagging places so that Max has knowledge of about 25 sights and 50 additional landmarks, which he uses as navigation points for his orientation.

In this scenario an episode is a single tour of a visitor accompanied by Max through the city. The visitor is supposed to explore the city and to look at sights of his interest. He can steer on his own using a Wii Remote controller or can ask Max to take control and guide him. After greeting the visitor, Max introduces himself and offers directly to give a guided tour or to just accompany the visitor in exploration of the town (an example dialogue is presented later on). New events are stored automatically in the memory and indexed along the six dimensions as Max and the visitor interact and move through the city together. So far, speech recognition is not provided for the visitor, he has to choose from a set of natural language sentences. Note that Max’ memory is not empty at the start. It already holds some episodes, which Max could have experienced before, but in this case are pre-programmed to equip Max with some initial memories as a starting point to guide.



Fig. 2. The scenario comprises the virtual agent Max standing together with the visitor on a floating hover disk in a CAVE-like virtual reality environment. Besides wandering through the city of Virtual Tübingen and looking at sights from “normal ground perspective” the city can also be examined from a bird’s eye view.

5 Episodic Memory in Interaction

For our scenario we identified four general interaction cases that make use of the episodic memory system.

An Accompanied Exploration. If the visitor chooses to explore the city on his own, Max just accompanies him (or her) and watches the visitor’s actions. At any point during the exploration Max can be asked about sights encountered. Max uses his memories to suggest where to go next. These recommendations are based on what sights have been visited so far in comparison to prior episodes. Figure 3 shows an on-going episode and two related completed episodes. Although Episode 2 has a slightly different path than the on-going episode it still may be a better reference where to go next than Episode 1 based on the content of the intentionality and causality indices.

A Guided Tour. If the visitor chooses to be guided, Max has the goal to give a tour and takes over the steering control. He remembers a “master” episode where he has given a tour before. If Max is not interrupted he will visit the same places and give the same information. On request by the visitor he may change the tour at a any point. He will then try to adapt to the visitor’s wishes with the help of different tours he has given before and with the knowledge he has. Consider Episode 1 as the master for the on-going episode (see Fig. 3). If Max is

interrupted at stop 2 and asked for a specific sight he might remember Episode 2 where the sight would be encountered next. That episode would become the master episode and at the end Max memorizes the new episode which is a mixture of Episode 1 and Episode 2.

A Self-Explanation. If the visitor asks Max why he has recommended a certain sight or why Max has executed a specific action, Max is able to explain himself. He does this by remembering the addressed event's causality and telling it to the visitor.

A Summary. If the visitor asks for a summary Max can recall all seen sights so far from his memory and tell them in the order of visits to the visitor.

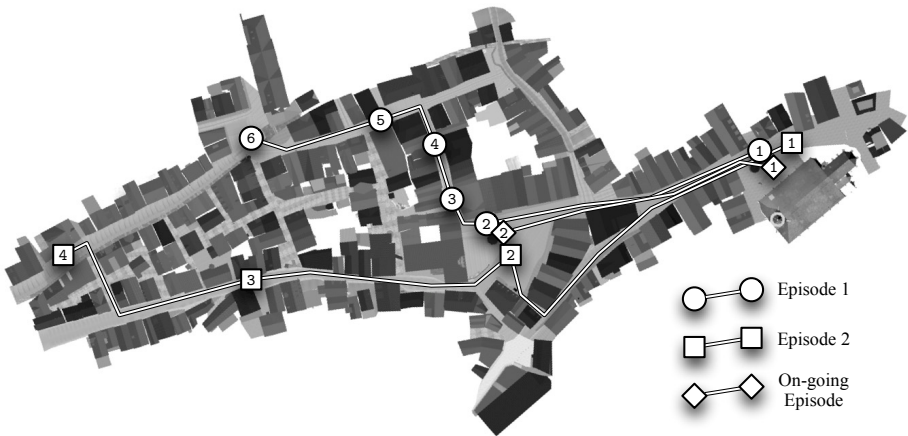


Fig. 3. Shown are two completed and one on-going episode visualized on the map of Virtual Tübingen

5.1 Excerpts From a Tour with Max Through Virtual Tübingen

We give some examples (translated to English) of how a visitor and Max interact during a tour in our through Virtual Tübingen. First Max greets a visitor and explains the navigation:

(MEM: Episode start)

MAX: Now, Felix, welcome to virtual Tübingen. This model resembles the nice old town of Tübingen. Right now, we are located at the marketplace.

(MEM: BEL Greeted)

(GOAL: Explain navigation in VR)

(INTENT: Find out if navigation in VR needs to be explained)

MAX: Do you know how to move about the place?

(Search MEM if CURRENT VISITOR was guided before: if YES remind CURRENT VISITOR of previous event, if NO explain navigation with Hover Disk)

VISITOR: No I don't. → (MEM: INTENT: Explain navigation)

MAX: Now, we are standing on the Hover disk together. By this we can move about the place. You can control the Hover disk by using the Wii-Mote yourself. Alternatively, I can control the Hover disk for you.

(MEM: BEL Navigation explained)

[etc.]

Then during the tour Max uses his memory to tell an interesting story:

MAX: This is the *Holzmarkt*.

(GOAL: Explain nearby sights)

(REMEMBER: Similar episode which has a strong emotional impact)

(INTENT: Tell interesting facts about Holzmarkt)

MAX: Over there you can see the old book store of *Johann Friedrich Cotta*, the most important German publisher of his time. He liked his small place right next to the market, and important authors like *Goethe*, *Schiller*, and *Hölderlin* to name only a few visited him in his shop regularly. But many foreign visitors could hardly believe that so many famous authors were coming to this place, since the market was in a really bad shape, e.g. the stairs of the *Stiftskirche* were ruins. Also the name *Holzmarkt* comes from the simple items that were sold here during that time, mostly fire wood and copper items.

VISITOR: Well, that is an interesting story. → (MEM: Positive feedback)

MAX: I can also tell you about the *Stiftskirche* while we are here. Do you want to hear about it?

Finally Max uses memory at the end of the tour to summarize which places have been visited so far and answers a question why a specific sight was visited:

VISITOR: Can you summarize what we have done so far?

(MEM: INTENT: Summarize tour on visitor request)

MAX: Ok, here we go: After I explained how you can move through Virtual Reality, we started out at the market place of Tübingen. I showed you the city hall and told you about the *Neptunbrunnen*. Afterwards we had a look at the astronomical clock and listened to the Glockenspiel of the city hall's clock. From there you went through the *Holzgasse* to the *Holzmarkt* and I showed you the *Cotta* house and the *Stiftskirche*. After it I lifted us up into the air and I gave you an overview of the whole town. Then we landed in front of the *Jazzkeller* and now we are looking at the *Krumme Brücke*.

(MEM: Summarized tour)

VISITOR: Why have we been at the *Jazzkeller*?

(REMEMBER: Event containing location *Jazzkeller*)

MAX: Because you asked me about it when we were up in the air.

VISITOR: Thank you for the great tour. Unfortunately I have to leave now, but I'll be back. See you.

MAX: Ok. If you like we can continue from here on then. Farewell!

(MEM: BEL Bid farewell)

(MEM: Episode end)

6 Discussion

We have introduced a cognitively motivated episodic memory system which is integrated in a BDI-based agent architecture and employed in a virtual tour guide. Our scenario explains how the interaction between the virtual guide and a human visitor benefits from the use of the episodic memory. The guide can give recommendations without needing a large database of previous interactions, as he gets more knowledgeable through interaction with visitors by memorizing a new episode every time giving a tour. An added benefit of enhancing interaction lies in possibility to have the guide explain himself to the visitor, which makes the guide more comprehensible and authentic. Evaluating our approach will be subject of future work.

While it is not novel to build a system which makes use of episodic memory, the strength of our approach lies in the six event indices and their independence of a particular domain. Since these indices have been identified to describe general situations, events other than from the tour guide scenario could be memorized. Further, the fact that the described memory system is integrated within a BDI agent architecture is not essential for its application. The general requirements for a system to use our episodic memory system for enhancing human computer interaction are knowledge of the context and an agent system's capability to recognize own actions and plans.

Acknowledgements. This work has been supported by the Deutsche Forschungsgemeinschaft (DFG) in the Center of Excellence Cognitive Interaction Technology (CITEC) at Bielefeld University. We gratefully acknowledge the MPI for Biological Cybernetics for providing us with the model of Virtual Tübingen [12].

References

1. Becker, C., Leßmann, N., Kopp, S., Wachsmuth, I.: Connecting feelings and thoughts - modeling the interaction of emotion and cognition in embodied agents. In: Fum, D., Del Missier, F., Stocco, A. (eds.) *Proceedings of the Seventh International Conference on Cognitive Modeling (ICCM 2006)*, Edizioni Goliardiche, Trieste, Italy, pp. 32–37 (2006)
2. Brom, C., Pešková, K., Lukavský, J.: What does your actor remember? Towards characters with a full episodic memory. In: Cavazza, M., Donikian, S. (eds.) *ICVS 2007*. LNCS, vol. 4871, pp. 89–101. Springer, Heidelberg (2007)

3. Ho, W.C., Dautenhahn, K.: Towards a narrative mind: The creation of coherent life stories for believable virtual agents. In: Prendinger, H., Lester, J.C., Ishizuka, M. (eds.) IVA 2008. LNCS (LNAI), vol. 5208, pp. 59–72. Springer, Heidelberg (2008)
4. Johnson, W.L.: Agents that learn to explain themselves. In: Hayes-Roth, B., Korf, R.E. (eds.) Proceedings of the 12th National Conference on Artificial Intelligence (AAAI-1994), vol. 2, pp. 1257–1263. AAAI Press, Menlo Park (1994)
5. Leßmann, N., Kopp, S., Wachsmuth, I.: Situated interaction with a virtual human – perception, action, and cognition. In: Rickheit, G., Wachsmuth, I. (eds.) Situated Communication, Trends in Linguistics, vol. 166, pp. 287–323. Mouton de Gruyter, Berlin, Germany (2006)
6. Nuxoll, A.M., Laird, J.E.: Extending cognitive architecture with episodic memory. In: Holte, R.C., Howe, A. (eds.) Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence, pp. 1560–1565. AAAI Press, Menlo Park (2007)
7. Pfeiffer, T., Liguda, C., Wachsmuth, I., Stein, S.: Living with a virtual agent: Seven years with an embodied conversational agent at the Heinz Nixdorf MuseumsForum. In: Barbieri, S., Scott, K., Ciolfi, L. (eds.) Proceedings of the International Conference Re-Thinking Technology in Museums 2011 – Emerging Experiences, pp. 121–131. thinkk creative & the University of Limerick, Limerick (2011)
8. Rabe, F., Wachsmuth, I.: Cognitively motivated episodic memory for a virtual guide. In: Filipe, J., Fred, A. (eds.) ICAART 2012 – Proceedings of the 4th International Conference on Agents and Artificial Intelligence. vol. 1, pp. 524–527. SciTePress, Vilamoura (2012)
9. Rabe, F., Wachsmuth, I.: An event metric and an episode metric for a virtual guide. In: Filipe, J., Fred, A. (eds.) ICAART 2013 – Proceedings of the 5th International Conference on Agents and Artificial Intelligence, vol. 2, pp. 543–546. SciTePress, Barcelona (2013)
10. Tecuci, D.G., Porter, B.W.: A generic memory module for events. In: Wilson, D.C., Sutcliffe, G.C.J. (eds.) Proceedings to the 20th Florida Artificial Intelligence Research Society Conference (FLAIRS-2007), pp. 152–157. AAAI Press, Menlo Park (2007)
11. Tulving, E.: Episodic and semantic memory. In: Tulving, E., Donaldson, W. (eds.) Organization of Memory, pp. 381–403. Academic Press, New York (1972)
12. van Veen, H.J.A.H.C., Distler, H.K., Braun, S.J., Bühlhoff, H.H.: Navigating through a virtual city: Using virtual reality technology to study human action and perception. *Future Generation Computer Systems* 14(3–4), 231–242 (1998)
13. Zacks, J.M., Speer, N.K., Swallow, K.M., Braver, T.S., Reynolds, J.R.: Event perception: A mind–brain perspective. *Psychological Bulletin* 133(2), 273–293 (2007)
14. Zacks, J.M., Tversky, B.: Event structure in perception and conception. *Psychological Bulletin* 127(1), 3–21 (2001)
15. Zwaan, R.A., Langston, M.C., Graesser, A.C.: The construction of situation models in narrative comprehension: An event-indexing model. *Psychological Science* 6(5), 292–297 (1995)

System of Generating Japanese Sound Symbolic Expressions Using Genetic Algorithm

Yuichiro Shimizu, Tetsuaki Nakamura, and Maki Sakamoto

The University of Electro-Communications

Abstract. Japanese has a large number of sound symbolic words, onomatopoeia, which associates between sounds and sensory experiences. According to previous studies, a quantification of relationship between phonemes and images enables to predict the images evoked by onomatopoeia and to estimate meanings of onomatopoeia. In this study, we applied the quantification method and developed a system for generating Japanese onomatopoeias using genetic algorithm (GA). Our method uses 90 SD scales for expressing various impressions and genes for genetic algorithm which denote each phonological symbol in Japanese. Through genetic algorithm, the system generates and proposes onomatopoeias appropriate for impressions inputted by users. From the evaluation of our system, impressions of onomatopoeias generated by our method were similar to inputted impressions to generate onomatopoeias.

1 Introduction

Against a classical notion in linguistics that speech sounds and meanings of words are independent, the existence of synesthetic associations between sounds and sensory experiences (sound symbolism) has been demonstrated over the decades, e.g., Jespersen (1922); Köhler (1929); Ramachandran & Hubbard (2001); Sapir (1929). For example, *mal/mil* and *buba/kiki* for round and sharp shapes in Sapir (1929) Ramachandran & Hubbard (2001), respectively.

Japanese is known to have a large number of sound symbolic words, namely onomatopoeia. Onomatopoeia is an expression which imitates the sounds associated with objects or actions they refer to. For example, using “Bang !” (“*ba-n*” in Japanese), you can describe a loud sound which occurs when someone shot a gun. Onomatopoeia is also used to express more abstract concepts such as emotional conditions and the ways things are done. For example, “*wa-ku wa-ku*” for happy feeling and “*scrub scrub*” (“*go-shi go-shi*” in Japanese) describes the way in which you brush your teeth.

Previous studies such as Hamano (1998) point out a systematic sound symbolic relationship between Japanese phonemes and meanings (e.g., the vowel /i/ is associated with the sharp image). Fujisawa et al. (2006) quantified the relationship between phonemes and images and they constructed a model to predict the images evoked by onomatopoeias. Based on the model proposed by Fujisawa et al. (2006), Shimizu & Sakamoto (2011) developed a system to estimate the images of input onomatopoeias by users as shown in Figure 1. To estimate

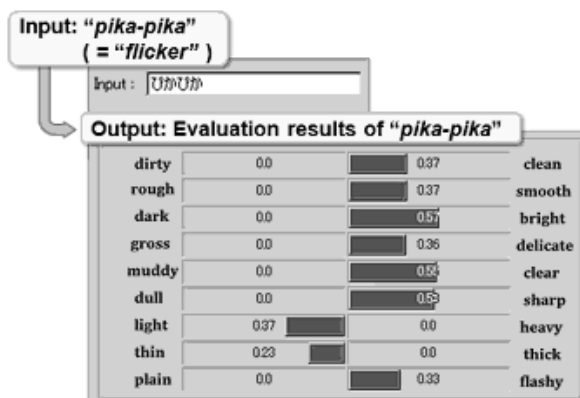


Fig. 1. A system for evaluating images of onomatopoeias

meanings of inputted onomatopoeias, the method uses the sound symbolic relationship between phonemes and meanings obtained through a psychological experiment. Meaning of each phoneme is calculated by making use of a Hayashi's mathematical quantification theory class I. The evaluation of the method showed that impressions estimated by the system were similar to those evaluated by humans.

In contrast to the system of Shimizu & Sakamoto (2011), the present study proposes a method which generates Japanese onomatopoeias corresponding to impressions inputted by users. In Japan, onomatopoeias are used frequently in comics and advertisements. Effective onomatopoeias in those fields are directly associated with sensuous experiences of readers or consumers, but it is very difficult to create such expressions. Thus, the technology which generates effective novel onomatopoeias corresponding to the impression specified by users is expected to be a technology which supports creators.

2 Method

2.1 Setting

Our system uses 90 SD scales as those expressing our impressions. These scales consist of scales which express various impressions: scales expressing impressions of haptic senses ("warm / cool", etc.), scales expressing impressions of visual senses ("blight / dark", etc.), scales expressing impressions of subjective senses ("clean / dirty", etc.), and so on. These scales are shown in Table 1.

Users of the system can decide the kinds of SD scales to be used to create onomatopoeias among 90 SD scales. The system uses the genetic algorithm to create onomatopoeias corresponding to inputted impressions.

In our method, created onomatopoeias are composed of 2 - 4 morae because Japanese has a lot of 2 - 4 morae onomatopoeias. We assume 7 components as

Table 1. Scales used in this study. These are used to generate onomatopoeias appropriate for impressions inputted by users.

warm / cool	smooth / rough	squeezing / not-squeezing
bright / dark	slippery / sticky	intolerable / tolerable
florid / plain	sharp / dull	oppressive / not-oppressive
soft / hard	thick / thin	twitching / not-twitching
lumpish / airy	firm / fragile	numbing / not-numbing
natural / artificial	regular / irregular	itching / not-itching
simple / complex	comfortable / uncomfortable	aching / not-aching
elegant / vulgar	easy / uneasy	stretching / not-stretching
static / dynamic	impressive / unimpressive	labored / not-labored
Western-style / Japanese-style	good / bad	pulled / not-pulled
modern / old-fashioned	intense / calm	severe / not-severe
masculine / feminine	repulsive / non-repulsive	stinging / not-stinging
sharp / mild	luxury / cheap	painful / not-painful
cheerful / gloomy	hot / cold	repulsive / not-repulsive
young / old	pressured / unpressured	fast / slow
pleasant / unpleasant	feeling-foreign-object / not-feeling	certain / uncertain
happy / sad	big / small	respondent / not-respondent
stable / unstable	heavy / light	indulgent / not-indulgent
individual / typical	momentary / continuous	comprehensible / incomprehensible
positive / negative	continual / continuous	visible / invisible
strong / weak	long / short	not-tired / tired
familiar / unfamiliar	wide / narrow	advanced / obsolete
like / dislike	deep / shallow	2-dimensional / 3-dimensional
fashionable / unfashionable	delicious / tasteless	direct / indirect
clean / dirty	sweet / not-sweet	intelligent / unintelligent
glossy / non-glossy	good-smell / bad-smell	strong (taste) / weak (taste)
wet / dry	reliable / unreliable	hot (taste) / not-hot (taste)
bumpy / flat	good-touch / bad-touch	bitter / not-bitter
elastic / non-elastic	good-quality / bad-quality	sour / not-sour
stretchable / unstretchable	gentle / strict	salty / not-salty

Table 2. Genes assumed in this study to form onomatopoeias

gene	phonological symbols
first consonants	E, k, s, t, n, h, m, y, r, w, g, z, d, b, p, ky, sy, ty, ny, hy, my, ry, gy, zy, dy, by, py
first consonants	a, i, u, e, o, N
special phonological symbols (middle)	E, N, Q, R
second consonants	E, k, s, t, n, h, m, y, r, w, g, z, d, b, p, ky, sy, ty, ny, hy, my, ry, gy, zy, dy, by, py
second vowels	E, a, i, u, e, o
special phonological symbols (tail)	E, N, Q, R, RI
repeating flag	F, T

those forming onomatopoeias and treat these components as genes for genetic algorithm. Each gene has plural phonological symbols as seen in Table 2. In the table, N, Q, R, RI, E, F, T respectively denote “the sound of the kana ‘n’”, “double consonant”, “lengthening of the previous phonological symbol”, “the

sound of the kana ‘ri’”, “this gene is not used”, “output onomatopoeia is used directly”, “output onomatopoeia is repeated twice”. Therefore, in our method, onomatopoeias are generated by combining phonological symbols assigned to these genes. The meanings of the seven genes are as follows (see Table 2.): (1) the first consonants (27 types of phonological symbols), (2) the first vowels (6 types of phonological symbols), (3) special phonological symbols which are inserted into middle of onomatopoeias (4 types of phonological symbols), (4) the second consonants (27 types of phonological symbols), (5) the second vowels (6 types of phonological symbols), (6) special phonological symbols of tails of onomatopoeias (5 types of phonological symbols) and (7) the repeating flag (whether the output onomatopoeia is used directly or repeated twice) (2 types of symbols).

Using these phonological elements, our system can generate Japanese kawaii expressions such as “pyokon-pyokon (py-o-E-k-o-N-py-o-E-k-o-N)” when the first consonant is “py”, the first vowel is “o”, the special symbol (middle) is “E”, the second consonant is “k”, the second vowel is “o”, the special symbol (tail) is “N”, and the repeating flag is T. “Pyokon-pyokon” is used to express a movement of something small or pretty.

2.2 Generating Onomatopoeias

The brief procedure of our method which generates onomatopoeias is as follows: (STEP 1) The system selects phonological symbols as genes at random. Then some onomatopoeias are generated by combining genes. (STEP 2) The onomatopoeias are evaluated by the quantification method developed in Shimizu & Sakamoto (2011). (STEP 3) The result of evaluation of generated onomatopoeias is compared to inputted impressions. Then some onomatopoeias similar to inputted impressions are selected. (STEP 4) Some onomatopoeias are generated by using genes (phonological symbols) of the selected onomatopoeias. (STEP 5) Some new onomatopoeias are generated based on mutation by fixed probability. (STEP 6) The procedure above (from STEP 2 to STEP 5) is repeated until onomatopoeias very similar to inputted impressions are generated or the procedure above is repeated until the fixed number of times.

The following is the detailed of the procedure. In Step1, initial individuals (onomatopoeias) are generated. The present system starts with 200 individuals. Namely, 200 individuals are generated in the first step. Each individual is generated by selecting one symbol for each gene in Table 2 at random and combining them.

In Step2, we used the calculation method which is a modified version of that in Shimizu & Sakamoto (2011). Impressions of generated individuals are calculated based on our quantification method. In this method, the impression of an onomatopoeia is composed by 90 SD scales mentioned at the front of this section. The system uses a database for category scores of phonological symbols, which stands for the sound symbolic relationship between phonemes and meanings. We got this relationship from a psychological experiment and quantified the relationship into category scores by the Hayashi’s mathematical quantification

theory class I. The database for category scores of phonological symbols shown in Table 2 (see Figure 3).

Therefore, each SD scale of the impression of an onomatopoeia x is calculated individually by Equation (1).

$$SD_i(x) = c1_i(x) + v1_i(x) + \overline{c2_i(x)} + \overline{v2_i(x)} + \overline{m_i(x)} + t_i(x) + r_i(x) + C_i \quad (1)$$

Where, in Equation (1), $SD_i(x)$ denotes the i -th SD value for x . $c1_i(x)$ denotes the i -th category score for the consonant of the first mora of x . $v1_i(x)$ denotes the i -th category score for the vowel of the first mora of x . $\overline{c2_i(x)}$ denotes the average value of the i -th category scores for the consonants of the second and later mora of x . $\overline{v2_i(x)}$ denotes the average value of the i -th category scores for the vowels of the second and later mora of x . $\overline{m_i(x)}$ denotes the average value of the i -th category scores for the special phonological symbols which are inserted into middle of x . $t_i(x)$ denotes the i -th category score for the special phonological symbol of the last mora of x . $r_i(x)$ denotes the i -th category score for whether x is an iteration structure. Finally, C_i denotes the value of the i -th constant term.

Table 3. The database for category scores of phonological symbols based on the Hayashi’s mathematical quantification theory class I. This database is consist of 90 category scores for each SD scale. In this table, “c1”, “v1”, “spsm”, “c2”, “v2”, “spst”, “repeat”, “const” respectively denote the first consonants, the first vowels: special phonological symbols which are inserted into middle of onomatopoeias, the second consonants, the second vowels, special phonological symbols of tails of onomatopoeias, the repeating flag, and the constant term.

SD	c1	v1	spsm	c2	v2	spst	repeat	const
	k ...	a ...	N ...	k ...	a ...	N ...	F T	
1st	0.23 ...	-0.02 ...	-0.86 ...	-0.51 ...	0.10 ...	-0.21 ...	0.32 -0.12	4.04
2nd	-0.29 ...	-0.16 ...	-0.28 ...	-0.45 ...	-0.07 ...	0.16 ...	-0.00 0.00	4.08
i -th
89 th	-0.05 ...	0.03 ...	0.07 ...	0.15 ...	-0.08 ...	-0.18 ...	0.16 -0.06	4.87
90 th	0.03 ...	0.07 ...	0.09 ...	0.18 ...	-0.05 ...	-0.10 ...	0.09 -0.04	4.83

In Step3, generated individuals are evaluated based on how much they are similar to the impression inputted by users. In our method, we treat an impression based on SD scales as a vector of SD values and use cosine similarity between two vectors as the degree of similarity between them. That is, the similarity between the impression of a generated individual (onomatopoeia) x and

that inputted by users is calculated by Equation (2).

$$s(\mathbf{v}(x), \mathbf{g}) = \frac{\mathbf{v}(x) \cdot \mathbf{g}}{|\mathbf{v}(x)| |\mathbf{g}|} \quad (2)$$

Where, in Equation (2), $\mathbf{v}(x)$ denotes the impression of x , \mathbf{g} denotes the inputted impression, and $s(\mathbf{v}(x), \mathbf{g})$ denotes the similarity between $\mathbf{v}(x)$ and \mathbf{g} .

In Step4, new generational individuals are generated based on the result of evaluation in Step 3. New individuals are generated based on selecting two old generational individuals and crossovering genes of these individuals. At first, two old individuals are selected according to the probability based on the similarity calculated in Step 3. Individuals having high similarity are selected easily. Next, genes of these individuals are split at random part and these split genes are crossovered so that new two individuals are generated. In our method, we use the one point crossover. This procedure is executed until the number of new generated individuals become the same as the number (that is, 200 in our system) of old individuals.

In Step5, the genes of new generated individuals are mutated based on fixed probability. That is, all of phonological symbols consisting of generated individuals are changed to different symbols which is not the present symbols based on fixed probability. In our method, because of shortness of a gene, this probability is very high (20%) so that the same individual might not be generated as much as possible.

In Step6, the procedure described above (from Step 1 to Step 5) is repeated until the similarity of an onomatopoeia which is the most similar to inputted impressions exceeds the threshold or the procedure is repeated until the fixed number of times. In our method, the threshold is 0.95 and the maximum number of repetition is 100000.

2.3 The Example of Execution

An example of output of our system is shown in Figure 2. The figure shows an example of output appropriate for the impressions. The upper figure is an example of SD values inputted by users and the lower one is an example of output appropriate for the impressions.

3 Evaluation

In the evaluation of our system, participants were asked to compare impressions inputted by SD scales with onomatopoeias generated by our system. We used various impressions for evaluation.

3.1 Participants

Participants are 15 Japanese males and females, aged 20-28 (the mean age is 22.9). They participated in this evaluation as volunteers.

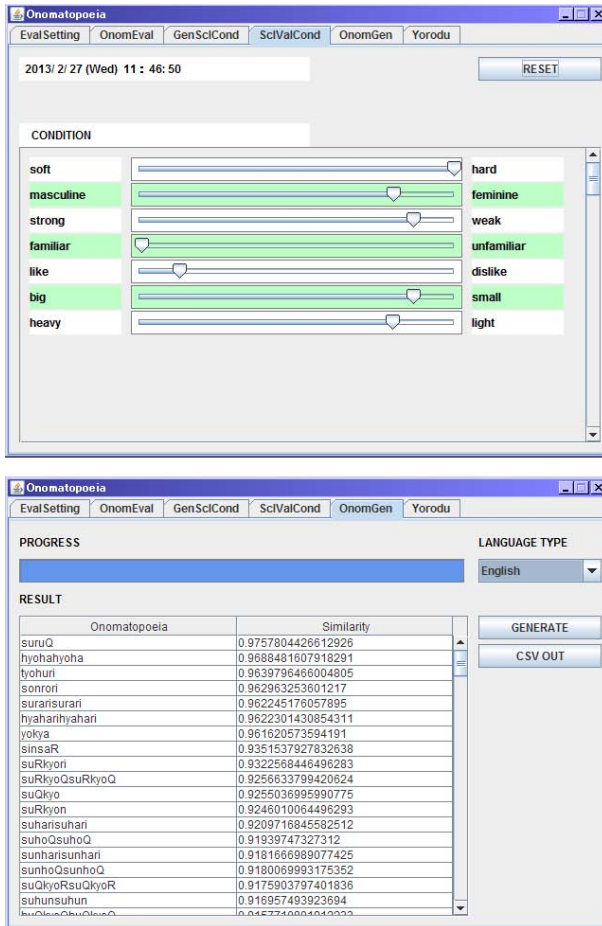


Fig. 2. The example of execution

3.2 Materials

We set 4 types of impressions to generate onomatopoeias for evaluation. The impressions are “cute”, “scary”, “romantic” and “lonely”. Participants decided the kinds of SD scales and inputted SD value so as to express the impression.

3.3 Procedure

It is difficult to evaluate our system because actual impressions evoked by onomatopoeias are subjective. So, we evaluate our system based on four items which participants evaluate how much generated onomatopoeias are consistent with inputted impressions. The items are as follows: (1) how much generated

onomatopoeias are coincident with inputted impressions, (2) how much novel generated onomatopoeias are, (3) how much interesting generated onomatopoeias are, and (4) how much generated onomatopoeias are easy to be understood.

Participants were asked to evaluate generated onomatopoeias based on 7-point semantic differential scales for these items. That is, (1) +3 to -3, (2) +3 to -3, (3) +3 to -3, and (4) +3 to -3.

3.4 Result

From the evaluation, the average values of the evaluated items answered by participants are shown in Table 4.

Table 4. The average values of the evaluation answered by participants.

“cute”		“scary”	
item	average value	item	average value
Coincidence	1.53	Coincidence	1.60
Novelty	2.20	Novelty	2.13
Interest	1.80	Interest	1.93
Understanding	1.20	Understanding	1.00

“romantic”		“lonely”	
item	average value	item	average value
Coincidence	-0.07	Coincidence	1.00
Novelty	2.27	Novelty	2.20
Interest	2.13	Interest	2.07
Understanding	0.67	Understanding	0.87

Considering these results, as a result of this evaluation, it is showed that impressions of onomatopoeias generated by our method were similar to inputted impressions to generate onomatopoeias. This result suggests that our proposal method is effective as a technology which supports creating novel onomatopoeias.

4 Conclusion

In this study, we developed a method which generates Japanese onomatopoeias corresponding to impressions inputted by users. This method is based on the genetic algorithm that phonological symbols treat as genes. As a result of the evaluation, it was showed that our method generated onomatopoeias which are similar to impressions inputted by users.

The present method generates only fixed-length onomatopoeias. Therefore, as a future work, we are planning to modify the method to control the length of generated onomatopoeias arbitrarily.

Acknowledgement. This work was supported by Grant-in-Aid for Scientific Research on Innovative Areas "Shitsukan" (No. 23135510) and (C) (No. 23500255) from MEXT, Japan.

References

- Jespersen, O.: *Language: Its Nature, Development and Origin*. George Allen & Unwin, London (1922)
- Köhler, W.: *Gestalt Psychology*. Liveright, New York (1929)
- Ramachandran, V., Hubbard, E.: Synaesthesia: A window into perception, thought and language. *Journal of Consciousness Studies* 8(12), 3–34 (2001)
- Sapir, E.: A study in phonetic symbolism. *Journal of Experimental Psychology* 12, 225–239 (1929)
- Hamano, S.: *The Sound-Symbolic System of Japanese*. Cambridge University Press, Cambridge (1998)
- Fujisawa, N., Obata, F., Takada, M., Iwamiya, S.: Impression of auditory imagery associated with Japanese 2-mora onomatopoeic representation. *Acoustical Science and Technology* 62(11), 774–783 (2006)
- Shimizu, Y., Sakamoto, M.: Japanese onomatopoeia generation system and image evaluation system by sound symbolism. In: *The 25th Annual Conference of the Japanese Society for Artificial Intelligence* (2011)

A Knowledge Elicitation Study for Collaborative Dialogue Strategies Used to Handle Uncertainties in Speech Communication While Using GIS

Hongmei Wang, Ava Gailliot, Douglas Hyden, and Ryan Lietzenmayer

Northern Kentucky University, Department of Computer Science,
Nunn Drive, Highland Heights, KY 41099, USA
(wangh1, gailliot1, hydend1, lietzenmar1)@nku.edu

Abstract. Existing speech enabled Geographical Information Systems (GIS) needs to have capabilities to handle uncertainties that are inherent in natural language communication. The system must have an appropriate knowledge base to hold such capabilities so that it can effectively handle various uncertainty problems in speech communication. The goal of this study is to collect knowledge about how humans use collaborative dialogues to solve various uncertainty problems while using GIS. This paper describes a knowledge elicitation study that we designed and conducted toward this goal. The knowledge collected can be used to develop the knowledge base of a speech enabled GIS or other speech based information systems.

Keywords: GIS, Knowledge elicitation study, Uncertainties, Human-GIS Communication, Collaborative dialogue strategies.

1 Introduction

Geographic Information Systems (GIS) are computer based mapping systems [1]. Since its early stage in the 1960s, they have been applied in various fields that use spatial data. Natural interfaces, e. g. speech and gesture enabled interfaces, have been proposed for GIS[2-6] over the years in order to further reduce the amount of training effort required from the GIS user. Some experimental GIS with natural interfaces have been developed. Early systems could accept only simple speech sentences which consisted of a few key words (such as CUBISON [7] and “Put-that-There” [8]) and/or simulated gestures (such as “Put-that-There” [8]). Along with advances in speech and gesture recognition in the computer technology field, some experimental natural interface-based GIS have been developed to accept more complicated speech input and/or pen-based gesture, such as QuickSet [9-11] and Sketch and Talk [12]. The most recent natural interface-based GIS can even recognize free-hand gesture, such as Dave_G [13].

Speech is used in most of the natural interfaces developed for GIS, and it is well known that natural language is not as precise as computer commands. The user’s speech requests can contain various uncertainties. They can be incomplete, ambiguous,

vague, and inconsistent [14-17]. Most of the existing speech enabled GIS do not handle such uncertainties well. They usually give a best guess only to the uncertain part in the user's speech request.

Human can usually successfully communicate with each other through collaborative dialogues, although their speech conversation also often contains various uncertainties. The success in human-human collaborative dialogues leads us to propose collaborative dialogues for speech enabled GIS to handle various uncertainty problems. The goal of this study is to collect human knowledge about how human communicators (human expert GIS operators in particular) handle various uncertainty problems in speech communication through collaborative dialogues.

To collect human knowledge, we need to apply some knowledge elicitation methods [18]. The observations and interviews are two of commonly used knowledge elicitation techniques [18-22]. They are direct methods of watching experts and interacting with them. In this study, we took these two techniques to collect human communicators' knowledge involved in handling various uncertainty problems while using GIS.

The first author's previous work also used these techniques on knowledge elicitation study for handling various uncertainty problems [23, 24]. However, the participant tasks designed in that study [23, 24] mainly focused on handling the vagueness problem. Each of these tasks involved speech communication of a vague spatial concept, *near*. The variety of uncertainty problems and their corresponding collaborative dialogue strategies that were discovered from that study were limited. The tasks designed for the participants to work on during the observation period in this study focused on various problems, instead of specifically on the vagueness problem.

2 Research Design

The research questions in the study include: (1) What kinds of uncertainty problems can occur in speech communication of spatial information requests while using GIS? (2) What collaborative dialogue strategies do human GIS operators take to handle these uncertainty problems? (3) How does a human GIS operator reason and make a decision during the process of communicating with the user, in particular, when the communication involves uncertainties?

2.1 Design of Participant Observation

The first technique used to collect human knowledge in this study is the participant observation. We planned to invite pairs of GIS experts and non-expert GIS users to work together on a set of tasks (see Table 1). We would observe their collaborative dialogues to reduce uncertainties in the communication, which would answer the research question 1 and 2.

Table 1. Eight User Tasks in Participant Observation

Task No.	Task Content	Common Contextual Information
1	<i>Get a Florida map</i> (At first, you want to see a map of Florida with basic information at first, such as state boundary, county boundary, cities, major roads etc.)	Traveling to Florida for Vacation: Imagine that you are planning to have a 2 months of vacation over Florida and not familiar with Florida. You have several requests.
2	<i>Get a map of flooded area in Florida</i> (imaging that you want to know the areas that usually flood during the summer)	
3	<i>Get an Ohio map</i> (Imagine that you want to know where is the Ohio county in Kentucky).	Living in Kentucky: Imagine that you are living in Northern Kentucky Area (here, in this study, it means the Kenton County and Campbell county). You have several requests.
4	<i>Get a map of Northern Kentucky Area</i> (imagine that you want to get a map showing basic information, such as county boundary, cities, roads etc).	
5	<i>Get a map of rivers in Northern Kentucky Area</i> (suppose that you are interested in only major rivers, not all streams).	
6	<i>Get a map showing 50-ft buffer zones around all rivers in Northern Kentucky Area.</i> (Suppose that you are interested in buffer zones around major rivers).	
7	<i>Get a map showing all daycares in Northern Kentucky area near your home</i> (Imagine that you have a child and need to select a daycare near your home (and work location if you work. If you work, suppose that your home and work location are close and can be considered as one destination.)	
8	<i>Get a map showing all KY cities near your "home city" Lexington</i> (imagine you are looking for jobs around home city)	

There are two sets of tasks (Table 1) for each pair of user participants to complete during the participant observation. The first set of tasks was situated in the context of planning for traveling to Florida for a two-month vacation. Task 1 was intended to

have them communicate missing information in the user request or on the map result because they may have different understandings on what should be displayed on the map. Task 2 involves communication of a general concept, the “flooded area”, which can refer to flooded area at different levels. We expect that this task would drive the participants to use collaborative dialogues to handle the generality problem.

The second set of tasks was set in the context of living in Kentucky. Task 3 was designed for the ambiguity problem because the word “Ohio” in the northern Kentucky area may mean two different things, the state, Ohio, and the county in Kentucky, Ohio. Task 4 had the same purpose as Task 1. Task 5 and Task 6 were also designed for the generality problem because the term “rivers” can refer to major rivers or all streams. Task 7 and Task 8 were both designed for the vagueness problem and involved communication of a vague spatial concept, near.. The vagueness problem is more complicated than the other types of uncertainty problems because it involves both context-dependency and fuzziness problems. However, the first author of this paper had conducted a knowledge elicitation study that specifically focused on the vagueness problem before [25]. Therefore, we did not design so many tasks to cover the various situations that the vagueness problem can occur.

2.2 Design of Follow-Up Interview

We planned to interview the participants after the observation. Each pair of participants would have some common questions related to the uncertainties observed in the participant observation. The common questions focused on the uncertainty problems that were observed during the participant observation in this study or previous studies [23, 24]. If a type of uncertainty problems happened in the participant observation, we would summarize how it is handled in the observation, and then ask the participant what other strategies can be applied. If it does not happen, we would explain the problem definition at first, and then ask the participant what strategies he/she would take if it happened. This part of interview was designed to collect more data for the research question 1 and 2. It would also be helpful to answer part of the question 3, such as how to identify each type of uncertainty problems.

The GIS operator participants would have additional questions, which were focused on their reasoning process underlying their collaborative communication with users during the observation. We would ask them to use one of the examples that happened in the observation process to explain how they made the decision of taking one of collaborative strategies available in their mind to handle the uncertainty problems. This part of interview was designed to collect data for the research question 3.

3 Data Collection

The data collection for the knowledge elicitation study was conducted at the first author’s office at Northern Kentucky University in summer 2011. The laptop at her office was installed with GIS software, ArcGIS Map 9.3, and some software for screen video recording and audio recording for the study. Four pairs of GIS operator

and user participants participated in the study. All of the GIS operators were the first author's students in her past GIS classes at Northern Kentucky University. They were either college students or had already had full time GIS jobs. The user participants were students at Northern Kentucky University from different majors. Seven of them were native English speakers. One of them was a foreign student and spoke in English fluently.

Same as the first author's previous knowledge elicitation studies [23-25], the data collection process consisted of three sessions: introduction session, participant observation session, and interview session. This section details the process.

3.1 Introduction Session

The consent forms were distributed and collected at first after the participants came to the experiment location. They were also asked to fill out a brief questionnaire about their background, including their gender, age, education major, GIS software use experiences etc.

Next, we gave a brief introduction of the study to each participant. We explained basic GIS concepts and the eight tasks (Table 1) to the user participant. The eight tasks would be completed by using some GIS functions. So, we demonstrated these functions to the GIS operator participant at first, and then asked the operator participant to practice using these GIS functions.

3.2 Observation Session

Each pair of participants started to collaborate on each of the eight tasks after they finished the introduction work. The entire collaboration process was recorded via two video cameras and screen recording software (Figure 1) while the investigators observed their collaboration.

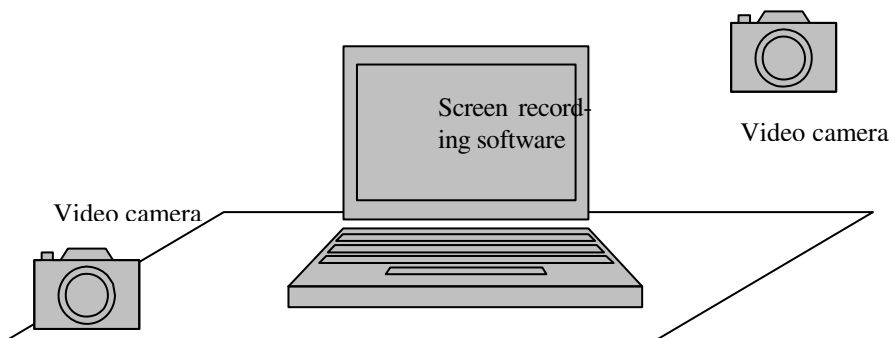


Fig. 1. Data Collection Setting for Participant Observation

The user participant initiated conversation with the operator participant for each task. The operator participant usually generated a map response for the user participant for each task. They usually need to use a few rounds of collaborative dialogues to complete each task.

3.3 Interview Session

The interview took place right after each pair of participants completed their tasks. The interview process was also recorded via multiple devices simultaneously (Figure 2). The investigators took written notes while interviewing the participant. Digital audio recording software on the computer and a digital video camera were also used to make sure that the interview process was recorded.

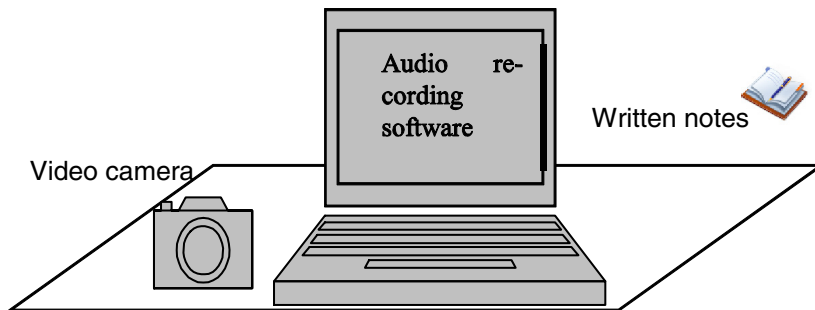


Fig. 2. Data Collection Setting for Interview

We conducted the interview by following the designed interview questions. The user participant was interviewed before the operator participant. This is because the questions for the user participant were less than those for the operator participant.

4 Findings

Due to the qualitative nature of the collected data, an interpretative reading method was applied to read and interpret all video, audio and written data. This yielded three major findings that correspond to the three research questions.

4.1 Uncertainty Problems

Several uncertainty problems were found in the speech communication while using GIS, including unclearness, forgetting details, incompleteness, vagueness, ambiguity, and generality. The first four types of uncertainty problems have already been discovered and explained in the first author's previous studies [23-25]. Therefore, we focus on explaining the last two types, that is, the ambiguity problem and the generality problem. Explanation to these two types of problems and some dialogue examples that were discovered from the participant observation are given in Table 2.

Table 2. Uncertainty Problems and Dialogue Examples

Type No.	Uncertainty Problem	Dialogue Example (O: GIS operator; U: User)
1	<i>Ambiguity:</i> Some terms in the user’s speech request have two different meanings.	U: ...Well can I get an Ohio map? O: An Ohio map? U: Mhmm, check and see the Ohio county that’s in Kentucky. O: The Ohio county? The county in the state of Ohio? U: Umm no, the Ohio county that’s in Kentucky.
2	<i>Generality:</i> A term in the user’s request corresponds to different GIS datasets that provide different levels of details.	U: Since we’re on Campbell county, can I see a map of the rivers around there? O: Okay. U: Well alright. And those are just the major rivers, right? O: Yes. U: No streams?

4.2 Collaborative Dialogue Strategies

The GIS operator usually has multiple collaborative strategies to handle each type of the uncertainty problems. Collaborative dialogue strategies for the ambiguity problem and the generality problem are given in Table 3.

Table 3. Collaborative Dialogue Strategies

Type No.	Uncertainty Problem	Collaborative dialogue strategies
1	Ambiguity	1. The GIS operator makes an assumption on the ambiguous part; 2. The GIS operator directly asks the user for clarification; 3. The user directly provides more context information to narrow down the options for the ambiguous part without being asked by the GIS operator
2	Generality	1. The GIS operator directly provides more detailed information at first and asks for clarification later; 2. The user assumes some context information, see the map results from the operator at first, and then clarifies the level of detailed information later if needed.

The results about collaborative dialogue strategies discovered in this study and previous studies [23-25] show that there are two common strategies for these various uncertainty problems. One common strategy, referred to as Strategy 1, is to show a

map result to the user based on the operator's assumption about the uncertainty in the user's request and then to wait for the user's correction if needed. The other common strategy, referred to as Strategy 2, is to ask the user to clarify the uncertain part in the user's request and then generate a map response based on the user's response to the user.

4.3 Reasoning Process

The GIS operator's reasoning process usually includes a few major steps: 1) Understanding the user request and interpreting it as the common goal of collaboration between the operator and the user; 2) Locating an appropriate GIS command for the user request; 3) Instantiating all parameters of the selected GIS command from the user request; 4) Executing the selected GIS command; 5) Returning responses to the user.

At step 3, uncertainty can arise if one of the parameters of a GIS command cannot be directly instantiated from the user's request. In this situation the operator must identify the uncertainty problem and make a decision about which collaborative dialogue strategy should be used to handle it. If the operator uses Strategy 1 and a parameter is instantiated based on the operator's assumption, the operator will need to wait for the user's correction feedback at Step 5. If the operator uses Strategy 2 and the uncertainty problem is asked to be clarified by the user, the operator returns a response to the user at Step 5 without uncertainties.

5 Conclusion

This paper describes the knowledge elicitation study that we conducted to help a speech enabled GIS to handle various uncertainty problems in human-GIS communication. The study results discovered two uncertainty problem types, ambiguity and generality, and their collaborative dialogue strategies, which are not shown in previous studies [23-25]. The paper also describes the preliminary findings about the operator's reasoning processing, in particular, about how to make a decision about what to do when the user's request contains some uncertainty.

These findings will be helpful for us to further improve our design of existing speech enabled GIS, in particular, design of the knowledge base and reasoning algorithms that are needed for the system to handle various uncertainties inherent in speech communication. These findings can also be extended for other speech based information systems.

Acknowledgments. This work is supported by Northern Kentucky University CINSAM grant (2011) and Northern Kentucky University's NSF Project FORCE summer UR-STEM Program (2011).

References

1. Chang, K.T.: Introduction to Geographic Information Systems. McGraw-Hill (2010)
2. Frank, A.U., Mark, D.M.: Language issues for GIS. In: Macguire, D., Goodchild, M.F., Rhind, D. (eds.) *Geographical Information Systems: Principles and Applications*, pp. 147–163. Wiley, New York (1991)
3. Mark, D.M., Svorou, S., Zubin, D.: Spatial terms and spatial concepts: Geographic, cognitive, and linguistic perspectives. In: *Proceedings of International Geographic Information Systems (IGIS) Symposium: The Research Adgenda*, vol. 2, pp. 101–112 (1987)
4. Florence, J., Hornsby, K., Egenhofer, M.J.: The GIS Wallboard: interactions with spatial information on large-scale display. In: Kraak, M.-J., Molenaar, M. (eds.) *Seventh International Symposium on Spatial Data Handling (SDH 1996)*, pp. 8A.1–8A.15 (1996)
5. Mark, D.M., Frank, A.U.: NCGIA Initiative 2, Languages of Spatial Relations. In: *Closing Report: National Center for Geographic Information and Analysis*, Santa Barbara, CA (1992)
6. Mark, D.M., Frank, A.U.: User interfaces for Geographic Information Systems: report on the specialist meeting. National Center for Geographic Information and Analysis (1992)
7. Neal, J.G., Thielman, C.Y., Funke, D.J., Byoun, J.S.: Multi-modal output composition for human-computer dialogues. In: Antonisse, H.J., Benoit, J.W., Silverman, B.G. (eds.) *Proceedings of the 1989 AI systems in Government Conference*, pp. 250–257 (1989)
8. 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*, pp. 262–270 (1980)
9. Cohen, P., Dalrymple, M., Moran, D., Pereira, F.: Synergistic use of direct manipulation and natural language. In: *CHI 1989*, pp. 227–234. ACM/ Addison Wesley (1989)
10. Oviatt, S.L.: Pen/voice: Complementary multimodal communication. In: *Proceeding of Speech Technology 1992*, pp. 238–241 (1992)
11. Oviatt, S.L.: Multimodal interfaces for dynamic interactive maps. In: *Proceedings of the Conference on Human Factors in Computing Systems (CHI 1996)*, pp. 95–102 (1996)
12. Egenhofer, M.J.: Multi-modal spatial querying. In: Kraak, J.M., Molenaar, M. (eds.) *Advances in GIS Research II (Proceedings of The Seventh International Symposium on Spatial Data Handling)*, pp. 785–799. Taylor & Francis, London (1996)
13. MacEachren, A.M., Cai, G., Sharma, R., Rauschert, I., Brewer, I., Bolelli, L., Shaparenko, B., Fuhrmann, S., Wang, H.: Enabling Collaborative GeoInformation Access and Decision-Making Through a Natural, Multimodal Interface. *International Journal of Geographical Information Science* 19, 293–317 (2005)
14. Wang, F.: Handling Grammatical Errors, Ambiguity and Impreciseness in GIS Natural Language Queries. *Transactions in GIS* 7, 103–121 (2003)
15. Bosc, P., Prade, H.: An introduction to fuzzy set and possibility theory based approaches to the treatment of uncertainty and imprecision in database management systems. In: *Proceedings of the 2nd Workshop on Uncertainty Management in Information Systems: From Needs to Solutions* (1993)
16. Owei, V.: An intelligent approach to handling imperfect information in concept-based natural language queries. *ACM Transactions on Information Systems* 20, 291–328 (2002)
17. Robinson, V.B., Thongs, D., Blaze, M.: Machine acquisition and representation of natural language concepts for geographic information retrieval. In: *Proceedings of 16th Annual Pittsburgh Conference*, vol. 1, pp. 161–166 (1985)
18. Cooke, N.J.: Varieties of knowledge elicitation techniques. *International Journal of Human-Computer Studies* 41, 801–849 (1994)

19. Boose, J.H., Bradshaw, J.M.: Expertise transfer and complex problems: using Aquinas as a knowledge-acquisition workbench for knowledge-based systems. *International Journal of Man-Machine Studies* 26, 3–28 (1987)
20. Cordingley, E.S.: Knowledge elicitation techniques for knowledge-based systems. In: Diaper, D. (ed.) *Knowledge elicitation: Principles, Techniques and applications*. Ellis Horwood Ltd, Chichester (1989)
21. Meyer, M.A., Paton, R.C.: Towards an analysis and classification of approaches to knowledge acquisition from examination of textual metaphor. *Knowledge Acquisition* 4, 347–369 (1992)
22. Olson, J.R., Biolsi, K.J.: Techniques for representing expert knowledge. In: Ericsson, K.A., Smith, J. (eds.) *Toward a General Theory of expertise*, pp. 240–285. Cambridge University Press, Cambridge (1991)
23. Wang, H.: Knowledge elicitation study towards development of a collaborative speech enabled GIS. In: *Intellectbase International Consortium Academic Conference* (2010)
24. Wang, H.: Knowledge elicitation: a case study towards development of a collaborative speech enabled GIS. *Journal of Applied Global Research* 4(17) (2011)
25. Wang, H.: A Knowledge Elicitation Study for a Speech Enabled GIS to Handle Vagueness in Communication. In: *The 14th International Conference on Human-Computer Interaction: Towards Mobile and Intelligent Interaction Environments*. III, pp. 338–345 (2011)

Part II
Gesture and Eye-Gaze
Based Interaction

Context-Based Bounding Volume Morphing in Pointing Gesture Application

Andreas Braun¹, Arthur Fischer², Alexander Marinc¹,
Carsten Stockl ow¹, and Martin Majewski²

¹ Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany
{andreas.braun,alexander.marinc,
carsten.stockloew}@igd.fraunhofer.de

² Technische Universit at, Darmstadt, Germany
{arthur.fischer,martin.majewski}@stud.tu-darmstadt.de

Abstract. In the last few years the number of intelligent systems has been growing rapidly and classical interaction devices like mouse and keyboard are replaced in some use cases. Novel, goal-based interaction systems, e.g. based on gesture and speech allow a natural control of various devices. However, these are prone to misinterpretation of the user's intention. In this work we present a method for supporting goal-based interaction using multimodal interaction systems. Combining speech and gesture we are able to compensate the insecurities of both interaction methods, thus improving intention recognition. Using a prototypical system we have proven the usability of such a system in a qualitative evaluation.

Keywords: Multimodal Interaction, Speech Recognition, Goal-based Interaction, Gesture Recognition.

1 Introduction

Smart environments are often comprised of a plethora of networked and user controllable devices. Those are typically controlled by various remote controls or combined systems providing simplified graphical user interfaces. Pointing at devices for manipulation is a natural form of interaction that is often performed unconsciously when using traditional remotes. It is possible to realize this pointing manipulation by using a virtual representation of the physical environment in combination with gesture recognizing sensors [1]. The straightforward approach of finding devices is using an intersection between pointing ray and bounding volumes of devices in the virtual realm [2]. However, if the controllable devices are small or occluded selection might become difficult or even impossible. In this case means have to be provided to allow selecting the devices. Various options are available, such as conflict resolution strategies, e.g. via menu selection [3], the usage of visual indicators for aiding selection [4], or - as it is used in this work - using contextual information to infer the intention of the user of interacting with a specific device. This work will present the following contributions:

- We propose a generic method to modify bounding volumes based on contextual information gathered by the environment or the interaction process
- We propose different methods of bounding volume morphing, such as static scaling, occlusion-based morphing and viewpoint-based space-filling methods [5].
- We test our method in a multimodal interaction scenario using a combination of speech and gesture

We use the contextual information generated by the smart environment to modify the selection process on a generic level by modifying the bounding volumes associated with the different devices, instead of modeling the uncertainty within the pointing process itself. By this generic approach we gain two distinct advantages, the contextual information allows to reduce the information required by other systems in multimodal interaction scenarios and the modification within the virtual representation allows other applications to directly use the modified bounding volumes. A particularly interesting application area for this method is multimodal interaction. Concerning gestural interaction a good candidate for an additional modality is speech. This allows interacting with devices by pointing at them and speaking out various commands. The intention as identified by Natural Language Processing applied to speech and the approximate can be considered context. E.g. if the user wants to make something “louder” this is unlikely to apply to lighting - if the user is pointing to the front he typically does not want to interact with devices behind him. Therefore if the devices are properly mapped to speech control it is possible to reduce the number of potential systems to interact with and use this information in the bounding volume modification. The overall process in this application scenario is following five steps; processing speech for interaction commands, modifying list of potential devices based on supported commands, modifying bounding volumes of candidate devices perform ray cast based on pointing direction and identify device and executing command on device.

2 Related Work

In the last few years novel interaction paradigms have seen a strong interest in the public eye. Particularly gesture interaction has seen considerable success; particularly in mobile applications with touch screens and gaming applications, with the Nintendo Wii and Microsoft Kinect.

There have been various research efforts to use gestural interaction in smart environments. Wilson et al have created the XWand, shown in Figure 1- left, a gesture interaction device based on accelerometers and infrared tracking of the device position [2]. The integrated sensors allow determining pointing direction and starting point, thus providing the ability to select modeled devices in a smart environment. The system also allows using speech commands to manipulate the selected devices. XWand models devices as Gaussian probability distribution, allowing for simple decision which device should be selected, however the method does not take into account ambiguous or occluded appliances. In our work we build upon a bounding

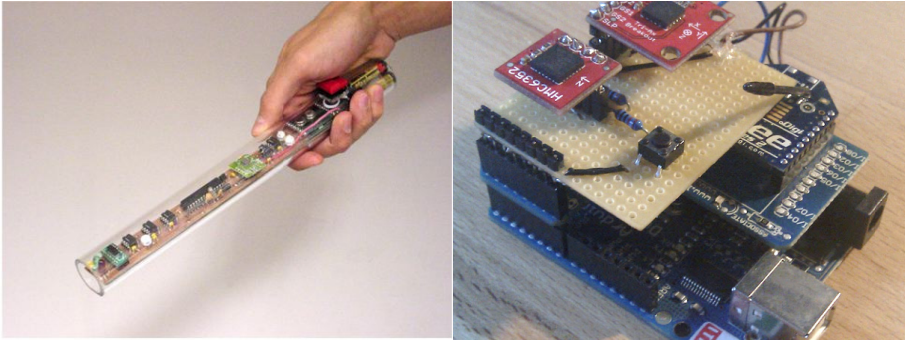


Fig. 1. Left - XWand gesture interaction device. Right - prototype interaction device.

volume approach previously presented [1] and introduce dynamically modified bounding boxes that change their shape based on the currently registered context, in this case speech and pointing direction. In contrast to the interaction device we have previously used (Figure 1- right) the new system is based on depth imaging.

Recognizing the intention of a person is a task typically performed subconsciously without rationalizing the motives of the conversation partner [6]. Even in simple conversations we evaluate the intentions continuously and use it as a supplement to our communication efforts to generate additional information that is important in the context of the conversation [7]. Heinze et al postulates that in inter-agent communication the recognition of intention is crucial if the transmission between the agents is flawed and ambiguous [6]. This is typically the case in Human-Machine-Interaction with natural input methods that mimic interpersonal communication [7].

3 Goal-Based Interaction in Context-Sensitive Smart Environments

The dynamic nature of an environment is making it difficult to distinguish between intentional interaction and random movements [8]. Goal-based interaction aims at abstracting explicit interaction from the user and instead of specific functions act

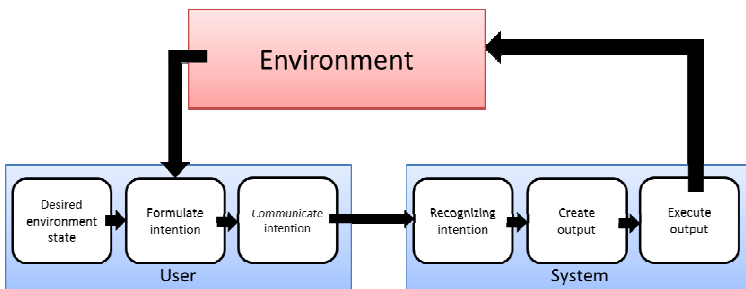


Fig. 2. Goal-based interaction without context support

based on the desired target of the interaction [9]. The general structure of a goal-based interaction system is displayed in Fig.2. A user is trying to achieve a desired environment state by formulating and communicating a specific intention. An interaction system is then trying to recognize this intention using the information communicated by the user. It will create the appropriate output and manipulate the environment accordingly.

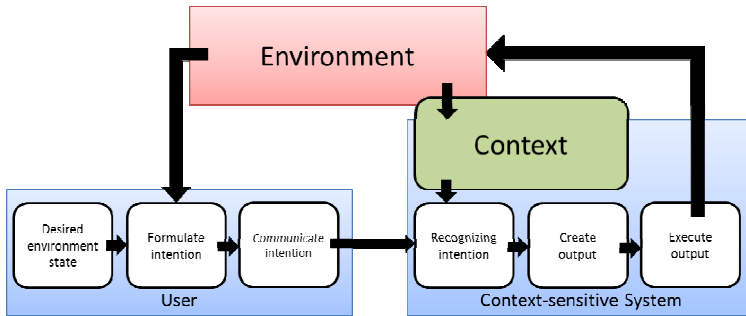


Fig. 3. Context-supported, goal-based interaction

This method however is not able to capture the implicit information. This is derived from interpersonal communication, wherein a considerable part of the information is exchanged implicitly within the current context; that is the situation surrounding the conversation and gives meaning to the specific interactions. In order to recognize this subtext it is necessary to monitor the user within the environment; analyzing the behavior and status to infer this information. The general structure of such a system is shown in Fig. 3 whereas the system has a second flow of information in order to recognize the intention using direct communication from the user and the context acquired in the environment. The latter method is particularly interesting concerning natural methods of interaction that abstract explicit functions from the user in order to allow interaction using the methods of interpersonal communication [10]. The question arises how we can use this concept in actual applications. A combination of speech and gesture is a common form of natural interaction that we are using to determine a suitable scenario for context-supported goal-based interaction. The direct channels of communication are the recognized gestures and the speech picked up by language processing. Combining these information channels with a modeled environment that is aware about its capabilities, those of the devices in the environment and activity information about the user we are able to create a scenario where we can improve the user experience by simplifying the interaction and making it more robust.

4 Bounding Volume Morphing and Multimodal Interaction

The combination of speech and gesture is a common form of multimodality [11, 12]. We use it in natural interaction, e.g. by pointing at a specific item, creating the implicit information that all subsequent information in this dialogue is centered on this item, without explicitly mentioning it every time. We can exploit this in a similar fashion for Human-Machine-Interaction. In this work we present a system supporting multimodal control of devices in smart environments. The supported method is the selection and manipulation of systems that are arbitrarily placed in the room. If the number of controllable devices is high it may be difficult to interact, e.g. considering small devices that have to be pointed at with gestural control, or numerous similarly named systems with speech control. If we combine both modalities we can create a model that supports and simplifies both methods of interaction by reducing the required inputs and increasing reliability. Based on this premise we have created a model that modifies the gestural selection process based on speech input and vice versa.

An overview of this process is given in Fig. 4. The user is communicating in a multimodal fashion using speech and gesture. The system is picking up this information and is additionally holding a model of the environment that is storing data about the different appliances, their capabilities and location. Both environment model and speech recognition influence the gesture recognizer while the final manipulation of the environment is depending on both speech and gesture recognition.

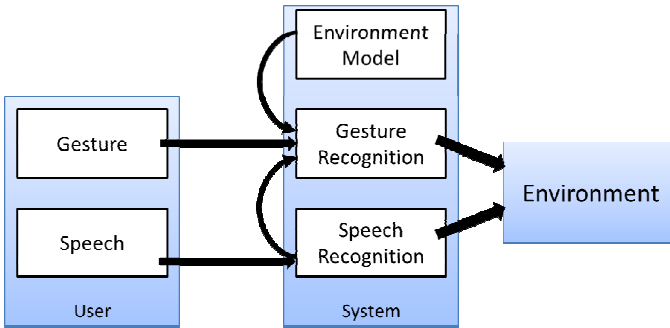


Fig. 4. Environment manipulation using speech and gesture

We are explaining this process by example of a user that is trying to control a lamp in a living room. He is pointing at the lamp he wants to turn brighter, however in the same region there are various other devices that make identification difficult for the gesture recognizer. Yet the system is aware of the device capabilities. The user now utters the words “brighter” indicating that he wants to control a device that is capable of changing lighting intensity. This information is going back to the gesture recognizer that discards devices that do not possess this ability, e.g. stereo or heating. The probability that the user is intending to select those devices can be lowered accordingly. One method to realize such a change in probability with regard to gesture recognition is modifying the bounding volumes of appliances, increasing or decreasing their

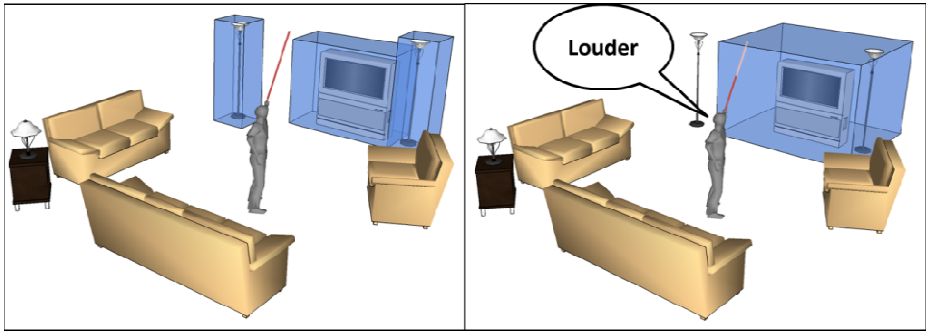


Fig. 5. Intersecting with a modified bounding volume after appropriate speech input

spatial representation in the environment model and thus adjusting the chances of intersecting this specific volume. To give an example, if there are three controllable devices, two lamps and a TV, and the user gives the command “louder”, the lamps can’t be affected lacking the capability. This behavior is shown in Fig. 5. If the lamps are discarded the bounding volume of the TV can be enlarged increasing the chances to be intersected.

The result is a two-step process, where first unsuitable appliances are discarded based on their capabilities and the results of the speech recognition and secondly the bounding volumes of all remaining devices are modified to increase the reliability of the gesture recognition.

Only modifying bounding volumes allows for generic application of various different methods. A first example is space-filling, whereas the bounding volumes are extended until they fill the available room; that is until they intersect the space boundaries or intersect with other bounding volumes. A second method is normalization, whereas the bounding volumes are extended to a fixed size, giving all objects the same probability of being intersected. Another example is uniform extension, leading to all bounding volumes being increased in size by a fixed ratio. All three methods are shown in Fig. 6 in a simple two-dimensional case.

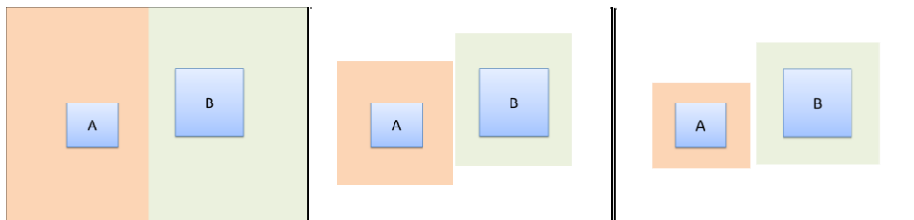


Fig. 6. Left space-filling method - middle, normalization method - right, fixed ratio method

When considering which method to choose it is crucial to think about the potential drawbacks of bounding volume based methods. We can distinguish two different types of errors. An occurring Type I error means that we are targeting at the actual

device but there is a bounding volume mismatch that does not allow us to properly select the system; Type II error means that an overly large bounding volume of another device is preventing us from being able to intersect the intended device [4]. Therefore it is crucial to select a method that is reducing both types of errors by creating optimal bounding volumes.

5 Prototype System

Based on the process described on the previous pages we have created a prototype system and installed it in our Living Lab. The devices in the lab are interfaced using a KNX bus system, that allows setting and manipulating various appliances within the premises, e.g. lighting, TV, windows and blinds. We have decided to use the Microsoft Kinect as gesture recognizing sensor using the OpenNI¹ framework. For speech recognition a dedicated microphone is used and interfaced with the CMU Sphinx framework² that allows recognizing speech commands using a combination of natural language processing with Hidden Markov Models. The virtual representation of the environment is based on X3D files, with the bounding volumes stored separately and modified accordingly. A software module combines the sensor input with the virtual representation and implements the device recognition using the bounding volume

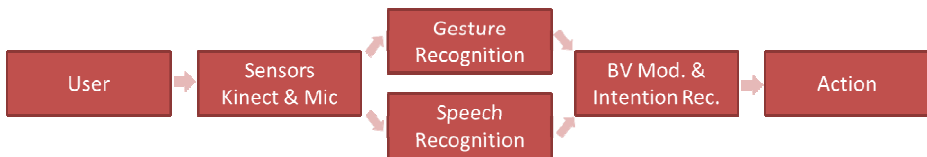


Fig. 7. Functional structure of the prototype system

modification methods presented previously. Afterwards this module sends the control signals to the KNX-networked devices. The overall structure of this prototype system is shown in Fig. 7. Given a set of possible devices and commands the system will combine them to determine the most probable device and execute the action intended with a command. For this purpose several cases in terms of the size of the sets have to be considered: In the trivial case one of the sets is empty and the system will just drop the current recognition process. In case there is only one possible device it will be assumed to be the final desired one and from all commands this device is capable of the most current command will be chosen. Finally, if the set contains multiple devices the most likely pair of device and command will be determined in four steps:

¹ <http://www.openni.org/>

² <http://cmusphinx.sourceforge.net/>

1. Remove all commands which are not part of the capability of any device
2. Remove all devices which are not capable of any of the remaining commands
3. Take the most recent command and increase the bounding volumes of all devices capable of it
4. Recalculate the intersection point of the pointing gesture and the environment.

The device the user is pointing on now is considered as being the users intended choice. Afterwards the final device-command pair will be forwarded and executed. In this procedure the third step defines that only the last command is a valid one in case of still existing uncertainty. This is due to the time frames around a detected pointing gesture. One or more commands arriving within one frame are expected to be corrections of the previous command. Changing step three to a sequential processing of all speech commands can be alternatively used. According to that corrections by the user would be realized by undoing previous commands instead of skipping the allegedly wrong commands.

6 Evaluation

We have performed a usability study in which the subjects had to perform simple tasks by using speech commands and pointing at the device to be controlled. The test was performed by nine users, aged between 21 and 29. Most had previous experience with gesture recognition systems, while most had little experience with speech recognition. The users had to perform a set of 11 different task controlling different devices in the environment, e.g. turning off lighting in the living room area. The devices were intentionally positioned to test cases that are relevant for context-based bounding volume adaptation, i.e. using small devices far away from the users and devices standing beside each other. The results were compared to a time-based selection, where interaction was enabled by holding a selection gesture for a certain amount of time. In this initial study we were mostly interested in getting an idea about the feasibility of our system and get on how users like the idea of using this multimodal interaction to control their smart environments. All subjects were able to perform all of the tasks with a noticeable learning effect from the first to the last tasks, reducing the number of wrong attempts and increasing the interaction time.

In a following interview the test persons considered the combination of speech and gesture to be preferable to gesture or speech alone. The subjects considered the interaction to be intuitive and easy to master and particularly liked how pointing can simplify the complexity of speech commands. However only one candidate could imagine using such a system right now to control devices and there were concerns about the performance of speech recognition, which can be attributed to the fact that the training had to be performed unspecified.

7 Conclusion and Future Work

We have presented a method that combines speech and gesture recognition to simplify interaction in smart environments. Using a virtual presentation of the environments we are able to control the gesture recognition using bounding volume modification. A test system based on the Microsoft Kinect and CMU Sphinx speech recognition was set up and tested with nine different subjects. The system compared favorably to time-based selection methods and all users were able to complete the defined set of tasks. Combining speech and gesture to control smart environments offers a huge potential. We can use the combined information to simplify interaction of the different modes. Using bounding volumes to realize this multimodal combination allows a direct integration in virtual representations of the smart environment and the possibility for modeling other aspects such as uncertainty or give an importance measure for the different devices, e.g. by changing the scaling factors based on confidence and a user-assigned weight. The initial results make us confident that the combination of speech and gesture to select and control devices is an approach that should be followed further.

We intend to upgrade our prototype system to a more capable speech recognition that does not require the user to hold a microphone, e.g. by using on-line speech recognition and microphone arrays. The gesture recognition performed favorably but can be improved using different feedback methods and a more precise skeleton tracker. In terms of bounding volumes we want to compare the results of different modification methods both quantitative in terms of how they fill the space and acquire a qualitative result on how they influence user experience. Another idea is to provide a measure how well-suited a given environment is for this kind of interaction based on size, capabilities and position of the included devices.

References

1. Braun, A., Kamieth, F.: Passive identification and control of arbitrary devices in smart environments. In: Jacko, J.A. (ed.) *Human-Computer Interaction, Part III, HCII 2011*. LNCS, vol. 6763, pp. 147–154. Springer, Heidelberg (2011)
2. Wilson, A., Shafer, S.: XWand: UI for intelligent spaces. In: *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pp. 545–552. ACM (2003)
3. Cao, X., Balakrishnan, R.: VisionWand. In: *Proceedings of the 16th Annual ACM Symposium on User Interface Software and Technology - UIST 2003*, pp. 173–182. ACM Press, New York (2003)
4. Majewski, M., Braun, A., Marinc, A., Kuijper, A.: Visual Support System for Selecting Reactive Elements in Intelligent Environments. In: *International Conference on Cyberworlds*, pp. 251–255 (2012)
5. Shneiderman, B.: Tree visualization with tree-maps: 2-d space-filling approach. *ACM Transactions on Graphics* 11, 92–99 (1992)
6. Heinze, C.: *Modelling intention recognition for intelligent agent systems* (2004)
7. Tahboub, K.A.: Intelligent Human-Machine Interaction Based on Dynamic Bayesian Networks Probabilistic Intention Recognition. *Journal of Intelligent and Robotic Systems* 45, 31–52 (2006)

8. Yamamoto, Y., Yoda, I., Sakaue, K.: Arm-pointing gesture interface using surrounded stereo cameras system. In: Proceedings of the 17th International Conference on Pattern Recognition, ICPR 2004, pp. 965–970. IEEE (2004)
9. Heider, T., Kirste, T.: Supporting goal-based interaction with dynamic intelligent environments. In: ECAI, pp. 596–602 (2002)
10. Valli, A.: The design of natural interaction. *Multimedia Tools and Applications* 38, 295–305 (2008)
11. Oviatt, S., Cohen, P., Wu, L., Duncan, L., Suhm, B., Bers, J., Holzman, T., Winograd, T., Landay, J., Larson, J., Ferro, D.: Designing the User Interface for Multimodal Speech and Pen-Based Gesture Applications: State-of-the-Art Systems and Future Research Directions. In: *Human-Computer Interaction*, vol. 15, pp. 263–322 (2000)
12. Quek, F., McNeill, D., Bryll, R., Duncan, S., Ma, X.-F., Kirbas, C., McCullough, K.E., Ansari, R.: Multimodal human discourse: gesture and speech. *ACM Transactions on Computer-Human Interaction* 9, 171–193 (2002)

Gesture vs. Gesticulation: A Test Protocol

Francesco Carrino^{1,2}, Antonio Ridi^{1,2}, Rolf Ingold²,
Omar Abou Khaled¹, and Elena Mugellini¹

¹ University of Applied Sciences Western Switzerland, Fribourg, Switzerland

² University of Fribourg, Fribourg, Switzerland

{francesco.carrino,antonio.ridi,omar.aboukhaled,
elena.mugellini}@hefr.ch, rolf.ingold@unifr.ch

Abstract. In the last years, gesture recognition has gained increased attention in Human-Computer Interaction community. However, gesture segmentation, which is one of the most challenging tasks in gesture recognition applications, is still an open issue. Gesture segmentation has two main objectives: first, detecting when a gesture begins and ends; second, recognizing whether a gesture is meant to be meaningful for the machine or is a non-command gesture (such as gesticulation). This paper proposes a novel test protocol for the evaluation of different techniques separating command gestures from non-command gestures. Finally, we show how we adapted adopted our test protocol to design a touchless, always available interaction system, in which the user communicates directly with the computer through a wearable and “intimate” interface based on electromyographic signals.

Keywords: Gesture segmentation, gesture interaction, test protocol, muscle-computer interface, system evaluation and interaction.

1 Introduction

For human beings, the gesture is a natural mean for effectively interacting with objects, tools and for communicating with other people. In 1980, the system developed by R. A. Bolt, known as “Put that there” [1] makes his entrance in the Human-Computer Interaction (HCI) field as the first system combining vocal and gesture interaction. After about thirty years, Microsoft launches Kinect going beyond existing devices such as Nintendo Wii or PlayStation Move, since it offers hands-free interaction. The success of this type of devices, also in research contexts, highlights the interest of exploiting gestures looking for novel interfaces and more natural ways of interaction with computers.

Previous works have explored input techniques based on a variety of input modalities: computer vision, special gloves, muscular activity, etc. However, independently from the approach used, the *gesture segmentation* remains an important problem to be addressed.

Within “gesture segmentation” we include two main aspects: detecting when a gesture is performed (beginning and ending) and therefore separate it from the previous

and the subsequent gestures; recognizing whether the subject is performing a command or a non-command gesture, and consequently changing the machine behavior. The typical example is the normal gesticulation that can be erroneously interpreted by the system as a valid command.

While detecting the beginning and the conclusion of a gesture is a problem often handled in literature, discriminating command and non-command gestures is an aspect rarely discussed. The problem is often avoided restraining the interaction to precise moments (synchronous systems) or using a limited set of unnatural gestures. Aiming at a more natural and always-available interaction experience, what is currently missing is a method that can aid evaluating different segmentation approaches, taking into account also the gesticulation.

In this paper we propose a novel test protocol focused on the evaluation of gesture segmentation approaches. In particular, the proposed method allows the differentiation between command and non-command gestures.

Finally, we show a first evaluation of our protocol in a test on the field. In particular, we adopted our test protocol during the design of an electromyography (EMG) based gesture interaction/segmentation approach. The main goal of the interaction system was to provide to the user an always-available human-computer interface based on subtle and motionless gestures. Triceps contractions were adopted to segment the gestures (i.e., performed gestures are recognized by the system as command-gestures only if they are performed during a contraction of the triceps).

This paper is structured as follow: the next chapter illustrates the works related to the domain of gesture segmentation and offers a survey of the different approaches used in recent studies; in section 3, we will show in detail the developed test protocol and how it has been adapted to the design of an EMG based interaction and segmentation system; finally, we will discuss the results and the possible test protocol improvements.

2 Background and Related Work

In literature, the segmentation problem has rarely been addressed independently, but rather within the gesture recognition context. Command gestures were often distinguished from gesticulation, setting aside very particular gestures for the interaction choosing command gesture too much wide and tiring and so too restrictive for an interaction claimed as “natural”.

When the gesticulation was considered in the context of interaction, this was often done aiming to deduct subject’s characteristics [2] or as a complement in voice-based interaction systems [3]. Alan Wexelblat [4] analyzed the gesticulation in order to deduct the most natural gestures and thus designing an interactive system to communicate naturally with virtual environments; the gestures were intended as command when used along with vocal orders.

The problem of separating continuous gestures has been addressed in particular contexts such as dance [5], theatre [6] and recognition of American Sign Language [7]. Kahol et al. [5] proposed a model called “Hierarchical Activity Segmentation to Represent the Human Anatomy” and they used low-level parameters to characterize motion in the various segments of the human body. Kelly et al. [8] presented a framework for continuous multimodal sign language recognition: they proposed a solution taking into account the epenthesis (i.e. the movement of the hand between the end point of the previous gesture and the start point of the next gesture).

However, these works did not consider the case in which the subject can move, interact with objects, and gesticulate, etc. between a gesture and the subsequent one.

Studies on always-available interfaces explored the problem of interacting with computer asynchronously during daily activities. Saponas et al. [9], in their EMG based interaction system, indicated the command gestures to the machine by clenching a fist: only when the fist was closed, the gestures performed with the other hand should be taken into account by the recognition system. However, in the following analysis they did not consider the influence of the segmentation system on the global system performance. Referring to similar experiences, Costanza et al. [10] tested their system in a particular scenario, providing a reasonably realistic environment (the subjects were walking through a predetermined path) maintaining enough experimental control to measure performance.

Actually, each work focusing on gesture recognition needs to develop its own approach to segment gestures. What currently lacks is an approach to evaluate the performances of the method used for the segmentation independently from the recognition and that could take into account non-command gestures.

In this paper we propose a test protocol that can be adopted to measure the performances of gesture segmentation approaches with regards to the gesticulation.

3 Test Protocol

In the first part of this section we will present an overview on the test protocol structure; subsequently, we will detail each step using as example our gesture recognition/segmentation approach for a real recognition/segmentation study presented in [11].

3.1 Protocol Overview

The test protocol (see Figure 1) is divided into four **phases** composed of eight different *steps*, concluded with one additional stage for the usability evaluation.

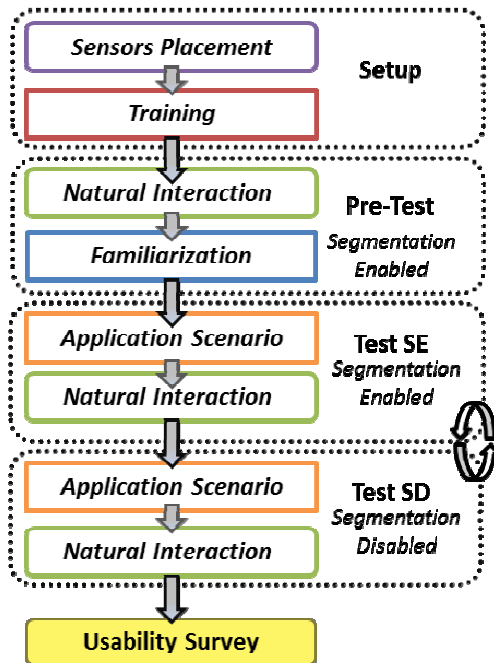


Fig. 1. Schematic representation of the test protocol structure

The first phase is the **setup** of the system. This part is required for two main reasons: the correct *sensors placement* (in the environment and/or on the subject) and *training* the system to adapt the recognition to the participants. This phase depends strictly on the approach and sensors used. For instance, cameras have different requirements than electrodes, and also the training is related with the classification strategy adopted. Once the setup is over, the segmentation system is activated and can detect when a command gesture is performed.

The **pre-test** session begins with the subjects performing the first *Natural Interaction* step. It consists in accomplishing predetermined tasks such as walking, manipulating as natural as possible objects with different characteristics, etc. Tasks and objects are closely related with the features used to distinguish the gestures. For example, during an approach based on muscular activity, “objects weight” and “handgrip type” are relevant features. On the opposite, during experiments involving computer vision techniques, the choice of objects according to features such as “color” or “shape” would be more suitable. This step precedes the *Familiarization* stage in which the subjects could finally understand and try the functioning of the system with the segmentation paradigm enabled. We chose to precede the *Familiarization* with a first *Natural Interaction* phase in order to evaluate whether or not knowing the interaction/segmentation approach could change the method of interaction.

After the *Familiarization* the **first test session** of tests begins. In this phase the segmentation is enabled and the subject performs one *Application Scenario* and one

Natural Interaction task. The *Application Scenario* is conceived to evaluate the accuracy of the system while the subjects interact with the system using gestures. The scenario has to deal with the challenging task to induce gesticulation in the user. The work presented in [4] gives some guidelines to follow in order to build a scenario enough involving to produce gesticulation.

In order to evaluate how the segmentation paradigm influences the recognition accuracy, the cognitive load and the system usability, a **second test session** is performed without segmentation.

The tests sessions (with and without segmentation) can be repeated several times in order to collect the desired amount of information.

Finally, the subjects conclude the experiment filling a *Usability Survey* based on System Usability Scale (SUS) [12].

3.2 EMG Test Protocol

We tested the protocol presented in the previous section in a real case [11]: evaluating the segmentation of two simple gestures of the wrist, segmented and recognized using EMG signals. Our test involved 11 participants, 8 men and 3 women, all of them were volunteers. They ranged in age between 23 and 30. Ten of the participants were right-handed and one was left-handed. The gestures were all performed with the preferred hand and none of the subjects had any known neuromuscular disease.

In order to use only not tiring gestures for the interaction, we adopted the wrist flexion and the wrist extension (see Figure 2). Keeping the palm of the hand in the sagittal plane, for the right-handed the wrist flexion corresponds to a left movement and the wrist extension to a right movement. For the left-handed the association is reversed.

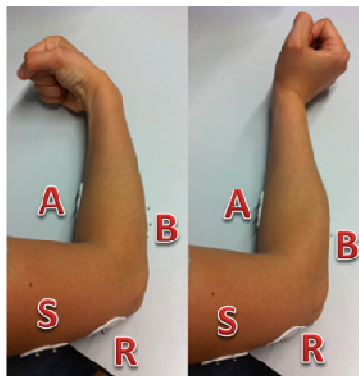


Fig. 2. Adopted gestures and electrodes placement: A and B detect wrist flexion and wrist extension; S is used for the segmentation; R is the reference electrode

The muscle used for the segmentation was the triceps. Contracting the triceps meant that the gesture performed at the same time had to be considered as a command gesture. All other movements, gestures and gesticulation had to be considered

non-command gestures. We choose the triceps contraction aiming at an almost-invisible, not tiring segmentation approach. These kinds of gestures have been defined “motionless gestures” [10].

Setup Phase

Sensors Placement. In our scenario, the system had to detect three gestures: the wrist flexion, the wrist extension, the triceps contraction. Therefore, the muscles directly involved in these movements were selected: for the wrist flexion the muscle designated was the *palmaris longus* and for the wrist extension the *extensor carpi ulnaris*. For the triceps contraction we monitored the *triceps brachii* activity. The skin was cleaned with alcohol according to SENIAM-recommendations [13], in order to assure a good skin-electrode contact. We did not shave or scratch the skin in order to evaluate a more interesting scenario for the HCI.

Training. A training phase was used in order to adapt the parameters of the used classifiers (Linear Discriminant Analysis) to each participant. We asked the subjects to perform the gestures as naturally as possible. A visual input indicated to the subject when to execute a specific gesture. A left arrow was used to indicate a left movement of the wrist, a right arrow for a right movement. The triceps contraction was associated with an up arrow. The training session lasted about one minute for each subject.

Pre-Test Phase

Natural Interaction. During the *Natural Interaction* step, the subjects were required to execute a series of tasks commonly done in the everyday life in order to quantify the false positives rate during these activities. The subjects were asked to:

- “Manipulate” five objects.
 - Displace the objects from a desk to another.
 - Putting the objects on the floor and then lifting them on the desk.
- Write and draw on a blackboard.

During the manipulation activities, the participants had to handle five different objects, with different weights. Also the grip changed but no indication on how to manipulate them was provided. The objects used were:

- Phone. The cell phone weighed about 150 grams. Given that it is a small object, many pretensions are possible.
- Bottle. The bottle was filled since the weight of 1 kg was reached. The material of the bottle was glass.
- Book. A thick book of 2 kg was used.
- A filled computer bag. It weighed 3 kg.
- A chair of about 7 kg.

The chair was chosen as border case to test the limits of our segmentation approach.

During the writing task, the subjects were asked to write a specific sentence on a blackboard using a chalk. The phrase was: “*It’s a rather rude gesture, but at least it’s clear what you mean*” (Katharine Hepburn).

The *Natural Interaction* phase ended drawing standard geometrical figures.

Familiarization. During this phase, the subjects received instructions about the functioning of the system and how to manage the *Application Scenario* in the following phases. In particular we explained how to use the triceps contraction to segment the other gestures.

The participants tested the system until they felt confident with the interaction. In our experiment this phase lasted in average 5 minutes.

Test Segmentation Enabled (SE) Phase

Application Scenario. In order to involve the user in the task and attempt to produce “natural” gesticulation, we asked each subject to interact with a slideshow presentation. In each slide there were the indication of the current slide and the indication to the slide they should reach.

To complete one tour, a total of 16 gestures, 8 *left* and 8 *right* were needed. In order to complete the *Application Scenario* two tours were required for a total of 32 gestures.

Each tour, during two slides, the subjects were asked to describe the pictures shown to them, evaluating if the images are good looking or not and arguing their reasons. This step was added to discern if the system detects false positives when the subject is gesticulating in a presentation context.

The participants were free to choose the duration of the interval between two interactions.

The *Application Scenario* was alternated with a *Natural Interaction* step.

Test Segmentation Disabled (SD) Phase

The whole first test phase was repeated a second time with the segmentation disabled (i.e., the input from the triceps were ignored) allowing to evaluate the impact of the segmentation system on usability, cognitive load and overall performances.

System Usability Scale

The SUS [12] was used in order to evaluate the system usability in terms of perceived complexity and difficulties, consistency and other aspects. The SUS consists of ten statements for which each participant specified the level of agreement on a 5 points scale.

Additional open questions were added to cover aspects missing in the SUS survey, such as muscular fatigue, and user opinion about strengths and weaknesses of the segmentation and interaction.

4 Discussion and Conclusion

Using the proposed test protocol to design an EMG-based interaction system [11], we could effectively collect different kinds of information (Figure 3) related to the selected segmentation approach: the *Natural Interaction* steps permitted to perform an accurate false positives analysis; the *Application Scenario* provided information about false negatives, accuracy rate and cognitive load; the *Usability survey* provided an important feedback about the system cumbersomeness and ergonomics. Finally, anticipating the first *Natural Interaction* phase before to the *Familiarization* phase allowed evaluating conscious or subconscious biases introduced when the subject knows the functioning of the system.

In our experimentation, the length of the experiment resulted in a good trade-off between completeness and fatigue, aspects particularly important when the muscular activity is monitored.

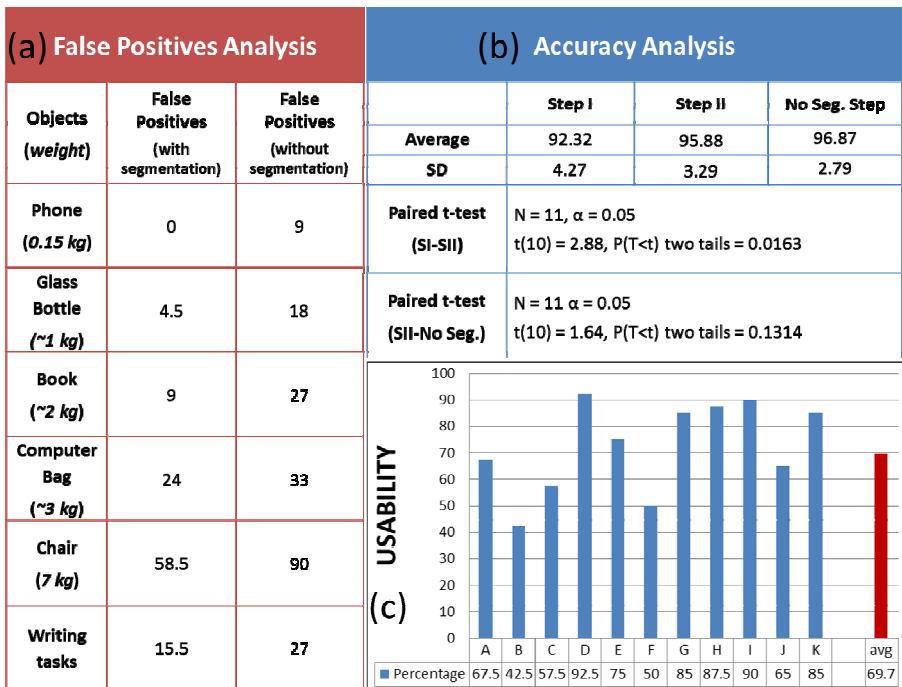


Fig. 3. (a) Results of false positives analysis, (b) accuracy analysis, and (c) usability analysis [11]

During the experiment, the *Natural Interaction* phase resulted in a uniform behavior among the subjects, with small variations in terms of seizing the objects and the energy/dexterity used in the write/draw tasks. On the contrary, the *Application Scenario* was visibly influenced by subjects’ cultural or behavioral aspects. Subjects not naturally inclined to gesticulate produced some movements only when strictly

required by the task. Different and maybe more stimulating scenarios, such as presented in [4], in which subjects were asked to describe in detail specific movies scenes, could improve the protocol.

From our experience, we learned that, in order to obtain significant results from the *Natural Interaction* phase, the choice of the features (e.g. weights and handgrip) and the consequent limit case (the chair) are of crucial importance.

In conclusion, in this paper we presented a novel test protocol for gesture recognition and segmentation methods. In particular our approach allows to easily separate command gesture from non-command gesture, such as gesticulation, providing an important tool to design new interaction systems. We showed a concrete example involving subtle gestures, for the interaction, and a motionless gesture, for the segmentation, demonstrating how the test protocol can be adopted to design a touchless interaction system.

Finally, in order to gain more relevance, our test protocol needs to be employed in similar researches comparing the results to estimate the need to include other overlooked gesture segmentation aspects.

References

1. 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, pp. 262–270. ACM (1980)
2. McNeill, D.: *Gesture and thought*. Continuum (2005)
3. Kettebekov, S.: Exploiting prosodic structuring of coverbal gesticulation on Multimodal interfaces. In: Proceedings of the 6th International Conference on Multimodal Interfaces, pp. 105–112 (2004)
4. Wexelblat, A.: An approach to natural gesture in virtual environments. *ACM Transactions on Computer-Human Interaction (TOCHI)* 2(3), 179–200 (1995)
5. Kahol, K., Rikakis, T.: Gesture segmentation in complex motion sequences. In: Proceedings 2003 International Conference on Image Processing (Cat. No.03CH37429), vol. 3, pp. II-105–II-108 (2003)
6. Billon, R., Nédélec, A., Tisseau, J.: Gesture recognition in flow based on PCA analysis using multiagent system. In: Proceedings of the 2008 International Conference in Advances on Computer Entertainment Technology - ACE 2008, vol. 139 (2008)
7. Li, Y., Chen, X., Tian, J., Zhang, X., Wang, K., Yang, J.: Automatic recognition of sign language subwords based on portable accelerometer and EMG sensors. In: International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multi-modal Interaction, vol. 17. ACM (2010)
8. Kelly, D., Reilly Delannoy, J., Mc Donald, J., Markham, C.: A framework for continuous multimodal sign language recognition. In: Proceedings of the 2009 International Conference on Multimodal Interfaces - ICMI-MLMI 2009, vol. 351 (2009)
9. Saponas, T.S., Tan, D.S., Morris, D., Balakrishnan, R., Turner, J., Landay, J.A.: Enabling always-available input with muscle-computer interfaces. In: Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology - UIST 2009, vol. 167 (2009)

10. Costanza, E., Inverso, S.A., Allen, R.: Toward subtle intimate interfaces for mobile devices using an EMG controller. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI 2005, vol. 481 (2005)
11. Carrino, F., Ridi, A., Mugellini, E., Khaled, O.A., Ingold, R.: Gesture Segmentation and Recognition with an EMG-Based Intimate Approach - An Accuracy and Usability Study. In: 2012 Sixth International Conference on Complex, Intelligent and Software Intensive Systems (CISIS), 4-6 July, pp. 544–551 (2012)
12. Brooke, J.: SUS-A quick and dirty usability scale, pp. 189–194 (1996)
13. Hermens, H.J., Freriks, B., Disselhorst-Klug, C., Rau, G.: Development of recommendations for SEMG sensors and sensor placement procedures. *Journal of Electromyography and Kinesiology: Official Journal of the International Society of Electrophysiological Kinesiology* 10(5), 361–374 (2000)

Functional Gestures for Human-Environment Interaction

Stefano Carrino^{1,2}, Maurizio Caon¹, Omar Abou Khaled¹,
Rolf Ingold², and Elena Mugellini¹

¹ University of Applied Sciences of Western Switzerland, Fribourg, Switzerland
{Stefano.Carrino,Maurizio.Caon,Omar.AbouKhaled,
Elena.Mugellini}@Hefr.ch

² University of Fribourg, Fribourg, Switzerland
{Stefano.Carrino,Rolf.Ingold}@unifr.ch

Abstract. In this paper, we describe an opportunistic model for human-environment interaction. Such model is conceived to adapt the expressivity of a small lexicon of gestures through the use of generic functional gestures lowering the cognitive load on the user and reducing the system complexity. An interactive entity is modeled as a finite-state machine. A functional gesture is defined as the semantic meaning of an event that triggers a state transition and not as the movement to be performed. An interaction scenario has been designed in order to evaluate the features of the proposed model and to investigate how its application can enhance a post-WIMP human-environment interaction.

Keywords: natural interaction, functional gestures, pervasive computing, human-computer interaction.

1 Introduction

This paper presents a model for the design of an opportunistic system for natural human-environment interaction. Natural interaction approaches aim to facilitate the control of technological devices through the use of the communication modalities typical of the human-human interaction [1]. Gestures, speech, gaze are few examples of typical modalities. Although natural paradigms have been conceived to improve learnability, several gesture-based applications 1) lack of expressivity (small lexicon) or 2) have a significant cognitive load for the user caused by the big number of gestures.

In this paper, we address these issues proposing an opportunistic context-aware model conceived to augment the expressivity of a small lexicon of gestures. The small size of the lexicon reduces the impact on the user cognitive load, whereas the opportunistic approach augments the vocabulary expressivity, with the results of an increased expressivity and a reduced cognitive load. A reduced number of gestures lowers the complexity of the system improving also the accuracy rate when using machine learning techniques. In fact, a classifier trained on a small number of gestures generally outperforms in terms of recognition accuracy the same system trained on a larger number of gestures [2]. Section 2 presents the works related to this project

in the field of gesture recognition and gesture vocabulary design. Section 3 defines the main concepts of the presented model. Section 4 details our model. Section 5 validates the proposed model discussing an interaction scenario. Finally, Section 6 discusses the achieved results.

2 Related Work

Several studies have investigated the definition and design of a gesture vocabulary for the natural interaction with objects, the environment and invisible computers (post-WIMP era). The definition of specific and generic gesture taxonomies is an important preliminary step designing the features of the interaction between the human and the machine. Researchers in HCI have proposed conceptual frameworks mixing gestures physical expression and semantics in the taxonomy (such as [3], [4] and [5]). For instance, in [3] Quek et al. define manipulative and semaphoric gestures. The first class involves “a tight relationship between the actual movements [...] with the entity being manipulated”; the second “any gesturing system that employs a stylized dictionary of static or dynamic hand or arm gestures”. These classes emphasize the relation of the gesture with the entity of the interaction and the signification of the gestures for the user.

Most of the studies on gesture interaction define an ad-hoc or rule-based gesture vocabulary to be used in the interaction (such for example in [6], [7] and [8]). Stern et al. [9] propose the definition of a gesture vocabulary based on psycho-physiological and technical factors for one-way (a human that commands a machine) communication. Differently from the approach we propose in this paper, the authors limit their study to the assumption one gesture – one command. Several studies take into account co-verbal gestures [3] [10], in order to study the relation between the gestures and speech or to extract multimodal information. In contrast, we focus on the single gestural modality leaving multimodal aspects to further studies. Researchers in the cognitive science domain are studying how object shapes can evocate functional and volumetric gestural knowledge [11]. In their work, Bub et al. define functional gestures as “gestures associated with the conventional uses of objects”. Starting from this definition, we extend it toward generic entities that can be real or virtual, associating the function evocation to the semantic meaning.

The gesture vocabulary design can be done following different approaches. Akers et al. [12] propose an observation-based design to reveal the optimal gestures for a given task. However, designing a gestures vocabulary implies taking into account many aspects. In fact, Prekopcsák et al. [13] identify four main design principles for everyday hand gesture interfaces in ubiquity, unobtrusiveness, adaptively and simplicity. Our approach takes into account these four aspects with a special attention to the adaptation parameter. Also Nielsen et al. [14] propose an interesting approach for developing gestural interfaces focusing on parameters as intuitiveness and ergonomics. The authors define some directives to follow in order to adhere to the most important principles in usability and ergonomics. They state that two are the possible approaches for the investigation of suitable gestures for HCI interfaces: bottom-up

and top-down. In particular, the bottom-up approach consists of taking the functions and finding the matching gestures. Our model affects this part of their procedure introducing the concept of functional gesture defined in the next section. Our procedure enhancement aims at providing a smaller gestures vocabulary in order to go beyond the learnability obtainable with standard approaches improving the recognition accuracy at the same time.

3 Definitions

3.1 Interactive Entities

Through gestures, users can interact with devices and tools that can be real or virtual. In this paper, we generically call these devices *interactive entities*. Our model classifies the interactive entities according to the following 2-elements taxonomy: *two-state entities* and *complex entities*. **Two-state entities** are simple entities that are characterized from having just the states ON and OFF. Lamps could typically belong to this category. **Complex entities** can be modeled by a finite-state machine representation with more than 2 states, in which each transition defines an action. It is convenient to distinguish the two-state entity class for the wide availability of devices that can fit this class and can be described with the same model.

On the other hand, the state machine representation of a complex entity is strictly linked to the functions of the device. Although automatic state-machine generation, configuration and deployment are not the focus of this paper research, solutions based on an ontological description of the interactive entities (such as presented in [15]) can help this process: ontologies can abstract heterogeneous devices as homogeneous resources.

3.2 Interaction Expressivity versus Cognitive Load

Combinations of multiple gestures to provide complex and rich commands to a system are a challenging mean of interaction. The effort required to learn or reproduce such language requires the users to make a remarkable effort. In common approaches, with the exclusion of sign language alphabets, it is rare to find gestural sentences composed of a concatenation of multiple gestures. And, consequently, the most diffused approach is to associate one gesture to one command. If such solution works well in systems and applications with reduced needs of expressivity, it can fall short to control complex interfaces (complex not complicate interfaces [16]).

On the other hand augmenting the vocabulary size is not always a viable solution. Previous studies assessed that a big number of commands can have a sensible impact on the user cognitive load and the associated information can be difficult to process. According to Miller's seminal work, seven "*plus or minus two*" appears to be the upper limit in the number of information that can still be processed with a not excessive load on the cognitive processing capacity of our brain [17]. Based on these results, we empirically limited the number of gestures for our interaction scenario to 8. These gestures are: select, turn on/off, next, previous, undo, increase, decrease and

exit (their functions and particularities are detailed in section 3.4 and Table 1). We can observe that there are not specifications or limitations on how to perform a gesture.

3.3 Functional Gestures

As mentioned before, we focus on gesture classification based on the function of a gesture and not the physical movements or posture. From this perspective, a gesture, or more in general a command, does not imply constraints about its physical realization, improving flexibility and user adaptation. From a more general, multimodal, point of view, a command can be provided using different communication channels. According to our classification, if a command function remains the same, the command belong to the same functional command class.

But what is a function? Representing an entity as a finite-state machine, a *function* is the semantic meaning of an event or condition that triggers a transition. An *action* is a specific transition. Examples of functions are select, next element, undo, etc. A *functional gesture* is linked to the concept of function instead of action (in conformity with the conscious gestures of semantic type described in [14]). The functional gestures are strictly connected to the functions of the entity that we are interacting with. For instance, the *next element* function has not meaning for a two-state entity.

The aim and the advantage of this approach is the abstraction between the physics of the gestures and its interpretation. A system configured to respond to functional gestures does not force the users to specific movements and can implement a one to many or many to one relation (i.e., one gesture for multiple actions).

In this research, we limit our lexicon to 8 functional gestures: select, turn on/off, next, previous, undo, increase, decrease and exit. Table 1 presents the 8 functional gestures integrated in our taxonomy; we emphasize in *italic* some of the main assumptions that should be taken into account designing an interface according to our model.

Table 1. Functional gestures: function-action association

Function	Action
Select	Select the <i>highlighted</i> element (for a specific entity in a certain state)
Turn on/off	Two actions: switch the highlighted entity condition: OFF ->ON or ON->OFF
Next	Highlight the next element in a <i>sorted</i> list
Previous	Highlight the previous element in a <i>sorted</i> list
Undo	Undo the last command
Increase	Increase the <i>main property</i> of the highlighted entity
Decrease	Decrease the <i>main property</i> of the highlighted entity
Exit	Exit from the current state

Highlighting implies a form of feedback to the user. It can be visual, acoustic, or multimodal.

Sorting implies the concept of order between the different elements.

Increase and decrease should be applied to a *main property* of an element. The degree of relevance of the property can change with the state of the entity. For example, interacting with a media center in the *play movie* state, the increase and decrease function can be associated to the loudness of the volume, whereas in the pictures browser slide-show they can be associated with the zoom level on the images.

Finally, the exit command implies to have a state-machine representation aware of the application interface (this can imply the need to memorize the historic of the interaction).

4 Model and Design Directives

The proposed model aims to enhance the interaction between the human and the environment finding a good balance between cognitive load and vocabulary expressiveness, in the context of gesture-based interaction. In smart environments the gesture interaction can be very varied. In order to address these challenging issues and focus on our research, we fixed some design directives.

4.1 Design Directives

We identified a number of rules and constraints that the gesture lexicon and the environmental interfaces of a smart home should respect in order to be modeled with the proposed approach.

Interaction lexicon should:

1. Have a moderate number of gestures to reduce the cognitive load for the user that have to recall the interaction to perform. E.g., seven more or less two is the range of numbers suggested by Miller in [17] and that we adopted in our scenario.
2. Define a set of semantic meanings, and not the cinematic and dynamic of the gesture itself that the designer can freely choose. Such meanings should be generic. Based on the research of Neßelrath et al. [18], in our scenario we propose 8 functions that we associated to the selected semantic meanings.
3. Be designed following the Nielsen et al. procedure [14] in order to adhere specific ergonomics and usability principles.

Environmental feedback interfaces should:

4. Be designed to be compatible with the generic meaning of the gesture vocabulary and increase the intuitiveness of the interaction. Section 4.4 discusses typical issues that should be taken into account designing the feedback for the user.

Functional gestures must be dynamically associated to precise *actions* on the entities present in the environment exploiting contextual information. In particular, we use two kinds of contextual information: the system state and the environmental data. The environmental data contain generic information such as lighting conditions, noise levels, user position and activity.

4.2 System as State Machine

The entity model we proposed is based on a finite-state machine representation of the devices present in the environment. Gestures are interpreted differently according to the current device state. Fig. 1 shows an example of state machine modeling a simple 5-states device. The figure illustrates an interaction with a media center system (used in the scenario presented in section 5).

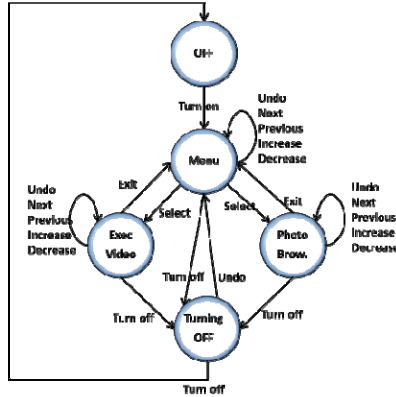


Fig. 1. Finite-state machine representation of a 5-state device (media center with video and photo browsing capabilities)

4.3 Contextual Information

Lighting conditions, noise levels, user position, user activity and the entity state are information that can be used in order to model the interaction properties from both interactive and technological perspectives.

From the point of view of the interaction, the context information can be used to propose the more appropriate gesture, modality or feedback (a seminal work on this subject is presented by Cheverst et al. [19]; we presented a prototype of context-based generation of multimodal feedback in [20]).

From a technological point of view, the contextual information can be used to improve the recognition task, for example selecting the most appropriate sensor in different conditions: low lighting conditions imply the use of sensors not based on RGB cameras, a highly noisy environment discourages the use of microphones or acoustic feedback, etc.

4.4 Feedback Design

In order to reduce the cognitive load on the user interacting with the environment the feedback design plays a crucial role. A highly dynamic environment should facilitate the interaction and help the user avoiding ambiguous interpretations of the system state. In fact, without an appropriate feedback the user can just assume, based on his

experience, the current state of the environment. Therefore, the interface should be able to adapt itself according to the user experience in the interaction. Novice users have greater needs of cues facilitating the interaction; on the other hand, expert users can feel such feedback as unnecessary, intrusive or distracting from the main task of the interaction [21]. The feedback design should take into account the following possibilities and needs:

- Visual (or multimodal) information suggesting the environment/entity state and the available interactions.
- Confirmation interface/state for important, long to undo actions. For example, the complete switch off of a device can imply a long time in order to return to a previous state (for instance, the turning off state in our scenario responds to this need).

5 Evaluation Scenario

We designed a scenario aiming at evaluating the features of our approach following the first two phases of the human-based experiment presented in [14]. We used this scenario in order to estimate the benefits and limits of the proposed approach. We use the user location (as proposed in [22]), the entity target of the interaction (as proposed in [23]) and its state as contextual information. The interactive entities are a media center, a radio, a lamp and a fan that can be remotely controlled (as depicted in Fig. 2). Our model uses the contextual information to dynamically adapt the link between the functional gestures and its related action.

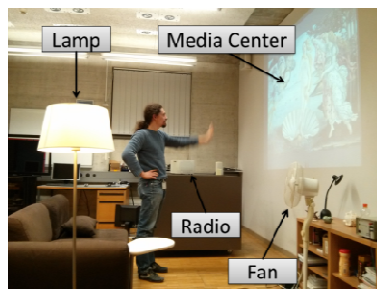


Fig. 2. Second scenario with a user and the interactive entities

In this section, we detail the interaction features related to the media center. For the other appliances we used a similar approach.

8 functional gestures are used in the interaction. The media center is modeled with a 5 finite-state machine: Off, Menu, Executing a video playlist, Photo browsing and Turning off (see Fig. 1). Table 2 details the dynamic, state-dependent links gesture-action. For instance, the functional gesture “turn on/off” is translated to the action “turn on” if the media center is in the Off state, “go to the turning off menu” if in the Menu, Executing a video playlist and Photo browser states, and “turn off” in the Turning off state.

Table 2. Functional gestures designed for a media center finite-state machine

Media center State Gesture	Off	Menu	Executing a video playlist	Photo Browsing	Turning off
Select	-	Select field	-	-	-
Turn on/off	Turn On	Turning Off menu	Turning Off menu	Turning Off menu	Turn off
Next	-	Next element	Next video	Next pic.	-
Previous	-	Previous element	Previous video	Previous Pic.	-
Undo	-	Undo	Undo	Undo	Undo
Increase	-	Volume Up	Volume Up	Zoom in	-
Decrease	-	Volume Down	Volume Down	Zoom out	-
Exit	-	-	Go to Menu	Go to Menu	-

In this example, 8 functional gestures are needed in the interaction. A standard approach with a relation one-to-one between the gestures and the actions needs 15 gestures (select, turn on, turn off, next element, previous element, next picture, previous pictures, next video, previous video, volume up, volume down, zoom in, zoom out, go to menu, undo).

Table 3 compares our approach with classical methodologies. Each entity is characterized by a finite state machine and a set of *actions*. The radio features 6 actions: turn on, turn off, next channel, previous channel, volume up, volume down. The fan and the lamp are modeled as 2-state entities.

Table 3. Number of gestures needed per device per approach. The functional approach is our contribution.

Interactive entities	Simple approach	Entity-aware	Functional
Media center	15	15	8
Radio	6	6	5
Lamp	2	2	1
Fan	2	2	2
Total number of gestures	25	15	8

In the first approach, we called “simple approach”, each device requires specific gestures and there is a mapping one-to-one between gestures and actions. The user should learn 25 gestures in order to fully interact with all the entities: 15 for the media center, 6 for the radio, 2 for the lamp and 2 for the fan. The “entity-aware” approach exploits the context information to recognize the target of the interaction but it is unaware of the entity state. This allows introducing 6 *functions* that model all the actions

for the radio, lamp and fan entities, and a subset of the media center actions. Therefore, the media center requires other 9 supplementary gestures to model the remaining actions, for a total of 15 gestures.

Finally, the third column presents our contribution: a state-aware system design that we called “functional” approach. In this case, the same functional gestures have different meanings according to the device state. With such approach, we can maintain the interaction expressivity limiting the number of gestures. In fact, the media center can be represented as finite-state machine (Fig. 1) and this allows defining 8 functional gestures that are enough to richly interact with all the entities in the environment as previously explained (Table 2). Such advantages are achievable only thanks to the abstraction work done at design time defining a specific state machine for each device. As mentioned before, ontology based approaches can help reducing this limitation.

6 Conclusion

The presented paper shows a natural interaction, context-aware approach that maximizes the lexicon expressivity of a limited set of gestures based on functions reducing the cognitive load on the user. In addition, a small number of gestures lowers the complexity of the system improving the accuracy rate of a classifier. The introduced 8-gesture vocabulary represents a generalized instance of our model and can be adapted to several different contexts for human-environment interaction in the post-WIMP era.

We can observe that in a domestic environment, interactive devices are typically simple entities that are easy to model and categorize. On the other hand, a bigger effort is required to model complex entities and the feedback to the user. For this reason we provide some design directives that in conjunction with the work of Nielsen et al. [14] constitutes a complete guideline for natural gestural interfaces design.

References

1. Carrino, S., Mugellini, E., Abou Khaled, O., Ingold, R.: ARAMIS: Toward a Hybrid Approach for Human-Environment Interaction. In: Jacko, J.A. (ed.) *Human-Computer Interaction, Part III, HCII 2011*. LNCS, vol. 6763, pp. 165–174. Springer, Heidelberg (2011)
2. Wachs, J.P., Kölsch, M., Stern, H., Edan, Y.: Vision-based hand-gesture applications. *Communications of the ACM* 54, 60 (2011)
3. Quek, F., McNeill, D., Bryll, R., Duncan, S., Ma, X.-F., Kirbas, C., McCullough, K.E., Ansari, R.: Multimodal human discourse: gesture and speech. *ACM Transactions on Computer-Human Interaction* 9(3), 171–193 (2002)
4. Karam, M., Schraefel, M.C.: A taxonomy of gestures in human computer interactions (2005)
5. Aigner, R., Wigdor, D., Benko, H., Haller, M., Lindlbauer, D., Ion, A., Zhao, S., Tzu, J., Valino, K.: Understanding Mid-Air Hand Gestures.: A Study of Human Preferences in Usage of Gesture Types for HCI, Microsoft Research Technical Report (2012)

6. Baudel, T., Beaudouin-Lafon, C.: Remote Control of Objects using FreeHand Gestures. *Communications of the ACM* 36(7), 28–35 (1993)
7. Kjeldsen, R., Hartman, J.: Design Issues for Vision- based Computer Interaction Systems. In: *Proc. of the Workshop on Perceptual User Interfaces*, Orlando, Florida, USA (2001)
8. Abe, K., Saito, H., Ozawa, S.: Virtual 3-D Interface System via Hand Motion Recognition from Two Cameras. *IEEE Trans. Systems, Man and Cybernetics, Part A* 32(4), 536–540 (2002)
9. Stern, H.I., Wachs, J.P., Edan, Y.: Optimal Hand Gesture Vocabulary Design Using Psycho-Physiological and Technical Factors. In: *7th International Conference on Automatic Face and Gesture Recognition (FGR 2006)*, pp. 257–262 (2006)
10. Kettebekov, S., Sharma, R.: Toward natural gesture/speech control of a large display. *Engineering for Human-Computer Interaction* 2254, 221–234 (2001)
11. Bub, D.N., Masson, M.E.J., Cree, G.S.: Evocation of functional and volumetric gestural knowledge by objects and words. *Cognition* 106(1), 27–58 (2008)
12. Akers, D.L.: Observation-based design methods for gestural user interfaces. In: *CHI 2007 Extended Abstracts on Human Factors in Computing Systems - CHI 2007*, pp. 1625–1628. ACM (2007)
13. Prekopcsák, Z., Halácsy, P., Gáspár-Papanek, C.: Design and development of an everyday hand gesture interface. In: *Proceedings of the 10th International Conference on Human Computer Interaction with Mobile Devices and Services - MobileHCI 2008*, p. 479 (2008)
14. Nielsen, M., Moeslund, T., Storing, M., Granum, E.: A procedure for developing intuitive and ergonomic gesture interfaces for HCI. In: *Proc. 5th Int. Workshop on Gesture and Sign Language based HCI, Genova, Italy* (2003)
15. Sommaruga, L., Formilli, T., Rizzo, N.: DomoML - an Integrating Devices Framework for Ambient Intelligence Solutions. In: *Proceedings of the 6th International Workshop on Enhanced Web Service Technologies - WEWST 2011*, pp. 9–15 (2011)
16. Norman, D.A.: *Living with complexity*. MIT Press (2010)
17. Miller, G.A.: The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *The Psychological Review* 63, 81–97 (1956), <http://www.musanim.com/miller1956>
18. Neßelrath, R., Lu, C., Schulz, C.H., Frey, J., Alexandersson, J.: A Gesture Based System for Context-Sensitive Interaction with Smart Homes. In: *Ambient Assisted Living, 4. AAL-Kongress*, pp. 209–219. Springer (2011)
19. Cheverst, K., Davies, N., Dix, A., Rodden, T.: Exploiting Context in HCI Design for Mobile Systems. In: *First Workshop on Human Computer Interaction for Mobile Devices*, pp. 1900–1901 (1998)
20. Perroud, D., Angelini, L., Khaled, O.A., Mugellini, E.: Context-Based Generation of Multimodal Feedbacks for Natural Interaction in Smart Environments. In: *The Second International Conference on Ambient Computing, Applications, Services and Technologies (AMBIENT 2012)*, Barcelona, Spain, September 23-28 (2012)
21. Bau, O., Mackay, W.E.: OctoPocus. In: *The 21st Annual ACM Symposium on User Interface Software and Technology - UIST 2008*, p. 37. ACM Press, New York (2008)
22. Greenberg, S., Marquardt, N., Ballendat, T., Diaz-Marino, R., Wang, M.: Proxemic interactions. *Interactions* 18(1), 42 (2011)
23. Carrino, S., Péclat, A., Mugellini, E., Abou Khaled, O., Ingold, R.: Humans and Smart Environments: A Novel Multimodal Interaction Approach. In: *Proceedings of the 13th International Conference on Multimodal Interfaces - ICMI 2011*, p. 105 (2011)

A Dynamic Fitting Room Based on Microsoft Kinect and Augmented Reality Technologies

Hsien-Tsung Chang, Yu-Wen Li, Huan-Ting Chen,
Shih-Yi Feng, and Tsung-Tien Chien

Department of Computer Science and Information Engineering, Chang Gung University,
Taoyuan, Taiwan
smallpig@widelab.org

Abstract. In recent years, more and more researchers try to make Microsoft Kinect and Augmented Reality (AR) into real lives. In this paper, we try to utilize both Kinect and AR to build a dynamic fitting room. We can automatically measure the clothes size of a user in popular brands or different country standards. A user can utilize gesture to select cloths for fitting. Our proposed system will project the video dynamically of dressing selected clothes in accordance with the captured video from Kinect. This system can be utilized in clothing store, e-commerce of clothes shopping, and at your home when you are confusing choosing a clothes to wear. This can greatly reduce the time you fitting clothes.

Keywords: Dynamic Fitting Room, Kinect, Augmented Reality.

1 Introduction

In recent years, Augmented Reality (AR) [1-6] is becoming an important and interesting technology for combine real live pictures and computed visualization images together. This can make user to interact between virtual and real worlds. In the beginning, AR is common used in entertainment, sport games, industry and even medical operation. And then it appears in the life enhancement applications, for example, it can be used in digital maps to demonstrate the navigation route into the real roads. It is interesting and really helpful to people.

Xbox360 is the second generation of Microsoft video game system and it was released on 22th November, 2005. Xbox360 gradually achieved market share from competitors with its successful hardware design and software supported. Especially the new controller/sensor Kinect was released on 4th November, 2010. The word Kinect is invented from the two words kinetics and connection. It utilized the VGA camera to capture the visible video from users and infrared camera to capture the distance between Kinect and users. After computation, Kinect can track the motion of two users and recognize 20 joints per user. Fig. 1 is the illustration of the 20 joints captured and calculated by Microsoft Kinect. Kinect[7-11] is recently utilized in many research areas.

How to make things convenient is an important issue to modern life. Therefore a lot of information technologies are invented to help modern people to achieve this goal. For example, Internet can speed up the transformation of information; mobile phone can communicate with others easier.

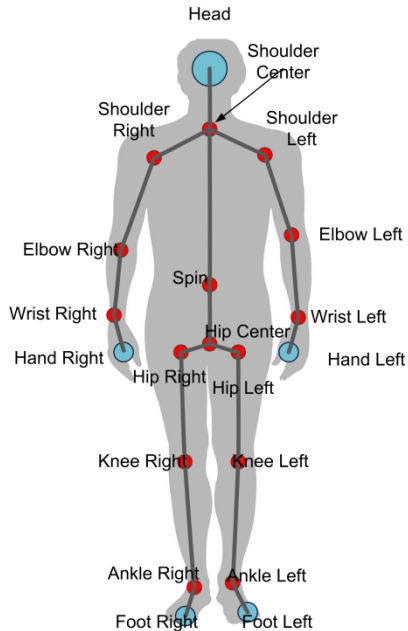


Fig. 1. Illustrated of 20 joints captured by Microsoft Kinect

Cisco blog posted ‘The future of consuming’ [12] on 25th July, 2011. It demonstrated a concept of video for a future fitting room. And Jade Jagger for Indiska Fashions uses the AR technology with marker to try on static clothes. This is the enlightenment idea to proposed this system.

In this paper, we design a dynamic fitting room which utilized the technologies of AR and Kinect. Users can use this system to try on clothes which is created in the digital wardrobe. And check the result of try immediately; even you move your body. You can change any clothes to dress yourself with AR technology to select the right style of for the coming party instead of putting on and taking off clothes. You will not sweat out and waste a lot precious time on fitting.

2 Dynamic Fitting Room

Our proposed dynamic fitting room is created for using on many scenarios. It can be placed in user’s house. And the existing clothes in the digital wardrobe are bought before. The user can try on any selected clothes in the digital wardrobe before taking the clothes out from the wardrobe in the real life. The dynamic fitting room can also be placed in a clothing store. When a customer enters the store, he/she can easily try

on different digital clothes sale in the store. It will save a lot of time on trying clothes on. The dynamic fitting room is also useful for e-commerce clothing stores, customers can see the real-time images that desired clothes trying on his/her own body. It will be more real than just watching the picture on models.

2.1 System Architecture

As demonstrated on Fig. 2, there two main sub-systems with Kinect in our proposed dynamic fitting room. One is called Wardrobe Screen, which displays the digital wardrobe and a user can select clothes in this screen. The other sub-system is called Dynamic Fitting Room, which will display the AR results with selected clothes.

These two sub-systems are run in different computer, and they communicate via network connection to exchange the needed information. For flexible, we store the digital clothes in a clothes database. The clothes database can be one’s private wardrobe, and it can also be an e-commerce clothing store, even it can be your friend’s wardrobe if your friend shares it.

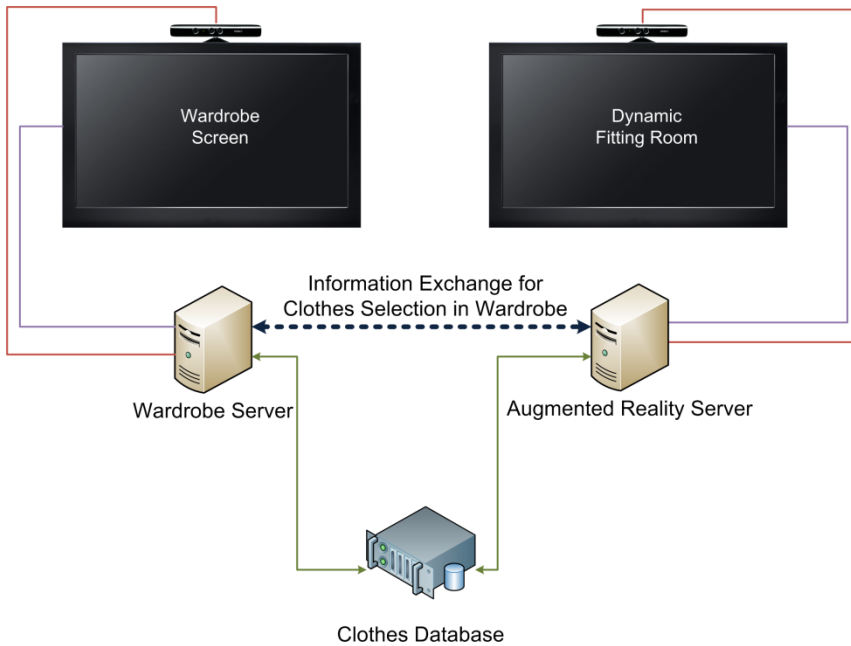


Fig. 2. System architecture of our proposed dynamic fitting room

Fig. 3 is the scenario that using our proposed dynamic fitting room. The user is standing between two large LCDs. The front side is placed the Dynamic Fitting Room sub-system, and the left side is the Wardrobe Screen sub-system. The user can first choose the clothes in the Wardrobe Screen sub-system in the left side and then check the AR results in the front Dynamic Fitting Room sub-system. When the user move his/her body, the AR result will display the real-time image on the screen.

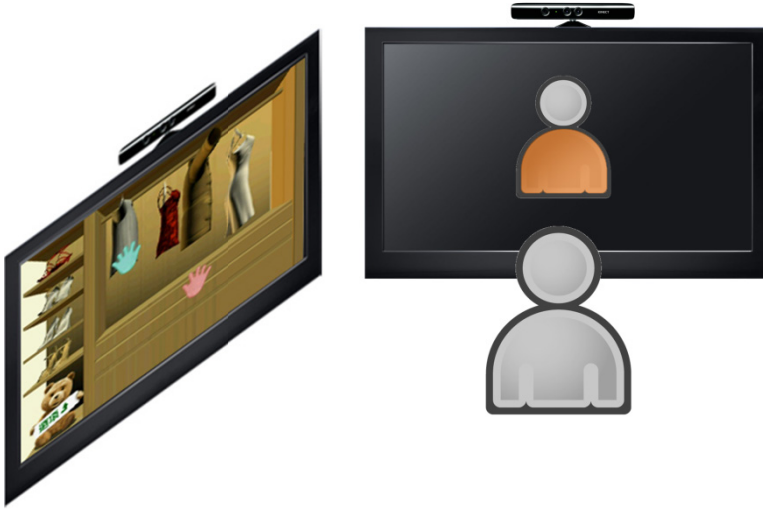


Fig. 3. The scenario of using dynamic fitting room

2.2 User Interface

Although the dynamic fitting room is composed with computers, it is impossible and weird to input information using a keyboard and mouse. The motion track function of Kinect is a good method for input information. Fig. 4 is the screen down of Wardrobe Screen sub-system. Two palms, blue and pink, represented two hands of the user. The left hand represented blue palm in the image can choose clothes by swiping left or right. The selected clothes will be bulged. The user can push right hand represented pink palm in the image to confirm your choice.



Fig. 4. Screen down of Wardrobe Screen sub-system

After choosing the clothes, the AR result is displayed in the front LCD. The user can move body to check the image with the selected clothes. In the same time, the Wardrobe Screen sub-system is temporally no function, because when you check the front LCD and move body may trigger some undesired function on the Wardrobe Screen sub-system. The user can raise two hands higher than head, and the Dynamic Fitting Room will be paused and the Wardrobe Screen sub-system will function again.

2.3 Fitting Clothes

We use Microsoft XNA Framework as our develop platform. The 3d model with skeleton of the digital clothes was designed in the Autodesk 3ds max. The concept of fitting clothes is receiving the joints and motion from Kinect. And then use the motion to trigger skeleton in the digital clothes. Fig. 5 displays the concept.

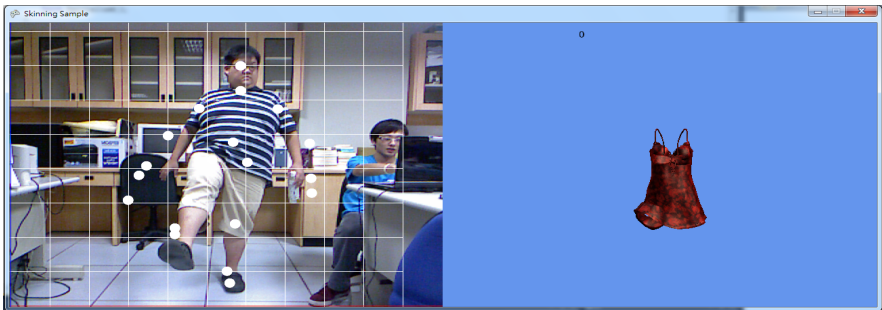


Fig. 5. Kinect joints trigger skeleton joints of 3Ds Max

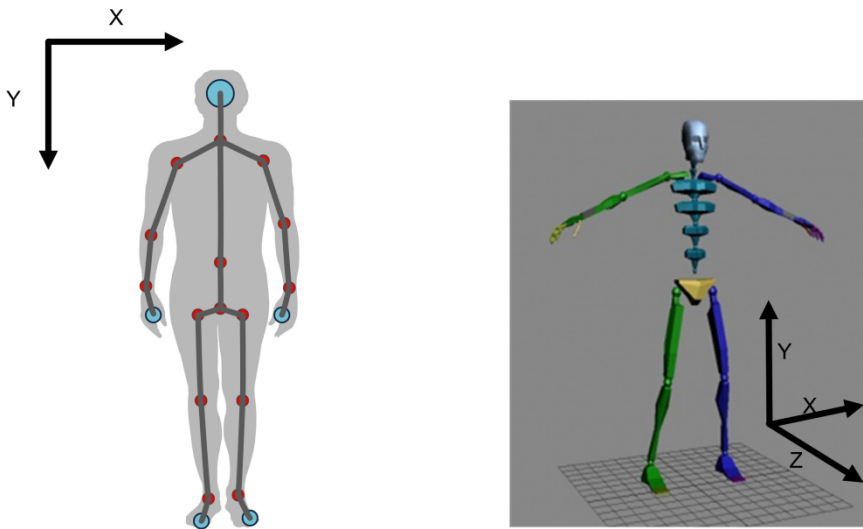


Fig. 6. The axis systems of Kinect SDK and 3Ds Max

Fig. 6 displays the axis systems of Kinect SDK and 3Ds Max. In the Kinect SDK, the axis information of each joint is represented as (x,y), and Kinect will also return a depth information of each joint. The depth information represents the distance between the joint and the Kinect inferred camera. It is not the real z-axis data. For example, if the user is standing straightly, the z-axis information of each joint must be the same. However, the depth information of the Hip Center and Head is different. The depth information needs proper conversion into z-axis information.

2.4 Size Issues

There are two main size issues in our proposed dynamic fitting room. What size you need is an important problem when you enter a clothing store. You can use your experience and try on some clothes and the possible size for the clothing store can be figure out. But it is not convenient. If you stand on our dynamic fitting room, the system will tell you or the seller what size you are. It will be a better solution than before. To solve this problem, we utilized the two Kinects in the front and left sides. The system will evaluate the user’s body height according to head/foot joints and the depth information using the front Kinect. And then calculate the possible waist length according to the oval perimeter using the width around the Hip Center joint from the front and left side Kinects. And last we calculate the area size of body according to the Kinect depth information from both Kinects. And then we use a rule-based method to determine the user’s size. Table 1-4 is the example rules for evaluate the user’s size. It can be adjusted for different clothing stores.

Table 1. Rule for body area(Pixel²) from front Kinect depth information

Lower bound	0	22000	30000	34000	36000	40000	42000
Upper bound	22000	30000	34000	36000	40000	42000	∞
Possible Size	NULL	XS~M	S~L	M~XL	L~3XL	XL~3XL	XL~4XL

Table 2. Rule for body area(Pixel²) from left Kinect depth information

Lower bound	0	12000	14000	18000	20000	23500	25000
Upper bound	12000	14000	18000	20000	23500	25000	∞
Possible Size	NULL	XS~S	S~M	S~L	M~XL	L~3XL	XL~4XL

Table 3. Rule for calculated body height(cm)

Lower bound	0	140	160	170	175	180	190	195
Upper bound	140	160	170	175	180	190	195	∞
Possible Size	NULL	XS~M	S~L	M~XL	M~2XL	L~3XL	XL~4XL	2XL~4XL

Table 4. Rule for calculated body waist(inches)

Lower bound	0	20	30	34	38	42	44
Upper bound	20	30	34	38	42	44	∞
Possible Size	NULL	XS~M	M~L	L~XL	XL~2XL	XL~3XL	2XL~4XL

The second size issue is what the size looks like to try on. Even we know the size we are, we want to try on different size according to difference style of clothes. We can create different 3D models for different sizes or just resize the 3D models. In our system, we just resize the 3D models for different Size.

3 Experiment and Results

The experiment is applied to 25 users. Before the size evaluation, we first ask user’s usual clothes size. And then evaluate by our system. Table 5 is the results of the 25 users. The results show that the evaluation of user’s size is quite closed to user’s claim. Fig. 7 is the demonstration that different users try on clothes. Fig. 8 is the size suggestion for different brands.

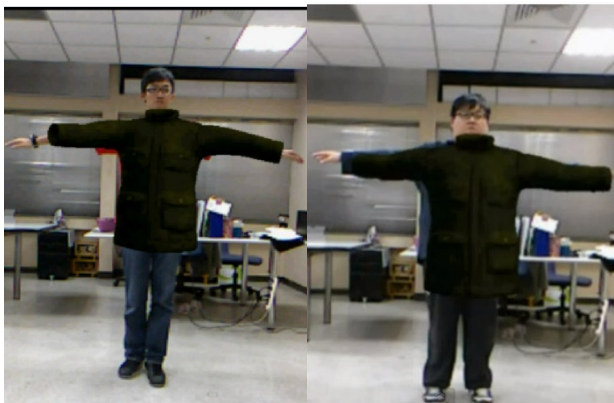


Fig. 7. Different users try on clothes

Table 5. Evaluate the size of user

Number	Claim Size	Result	Number	Claim Size	Result
01	XL	L	14	M/L	L
02	M	M	15	M	M
03	XL/2XL	2XL	16	L	L
04	3XL	3XL	17	2XL	3XL
05	M	M	18	M/L	L
06	M/L	L	19	M	M
07	S/M	M	20	M/L	M
08	M/L	M	21	XL	2XL
09	L	XL	22	M	L
10	L	XL	23	S/M	M
11	M	M	24	S/M	M
12	S/M	M	25	L	XL
13	M	M			



Fig. 8. Size suggestion for different brands

4 Conclusions

In this paper, we proposed a dynamic fitting room system utilized the Microsoft Kinect and Augmented Reality technologies. The system can show the real-time images that try on different digital clothes, and it also can evaluate user's clothes size. According to the experiment result, the evaluation of clothes size is quite closed to user's claim. This system can be utilized in clothing store, e-commerce of clothes shopping, and at your home when you are confusing choosing a clothes to wear. This can greatly reduce the time you fitting clothes.

Acknowledgement. The financial support by the National Science Council, Republic of China, through Grant NSC 101-2221-E-182-031- of the Chang Gung University is gratefully acknowledged.

References

1. Dingli, A., Seychell, D.: Blending Augmented Reality with Real World Scenarios Using Mobile Devices. *Technologies and Protocols for the Future of Internet Design* 258 (2012)
2. Graham, M., Zook, M., Boulton, A.: Augmented reality in urban places: contested content and the duplicity of code. *Transactions of the Institute of British Geographers* (2012)
3. Härmä, A., et al.: Techniques and applications of wearable augmented reality audio. In: *Proc. AES*, vol. 114 (2012)
4. Hondori, H.M., et al.: A Spatial Augmented Reality Rehab System for Post-Stroke Hand Rehabilitation. In: *Conference on Medicine Meets Virtual Reality, NextMed/MMVR20* (2013)
5. Huang, C.H., et al.: A CT-ultrasound-coregistered augmented reality enhanced image-guided surgery system and its preliminary study on brain-shift estimation. *Journal of Instrumentation* 7(08), P08016 (2012)
6. Yuen, S.C.-Y., Yaoyuneyong, G., Johnson, E.: *Augmented Reality and Education: Applications and Potentials*. In: *Reshaping Learning*, pp. 385–414. Springer, Berlin (2013)
7. Galatas, G., Potamianos, G., Makedon, F.: Audio-Visual Speech Recognition Incorporating Facial Depth Information Captured by the Kinect (2012)
8. Piyush, K., Stone, P.: A low cost ground truth detection system for RoboCup using the Kinect. In: *RoboCup 2011: Robot Soccer World Cup XV*, pp. 515–527 (2012)
9. Khoshelham, K., Elberink, S.O.: Accuracy and resolution of kinect depth data for indoor mapping applications. *Sensors* 12(2), 1437–1454 (2012)
10. Ono, M., et al.: SU-EI-91: Development of a Compact Radiographic Simulator Using Microsoft Kinect. *Medical Physics* 39(6), 36–46 (2012)
11. Wang, X.L., et al.: The Kinect as an interventional tracking system. In: *SPIE Medical Imaging*. International Society for Optics and Photonics (2012)
12. Veale, D.: *The Future of Consuming*, Cisco (2011), <http://blogs.cisco.com/tag/virtual-dressing-room/>

Gesture-Based Applications for Elderly People

Weiqin Chen

Department of Computer Science, Oslo and Akershus University College of Applied Sciences,
Postboks 4 St. Olavs plass 0130 Oslo, Norway
Weiqin.Chen@hioa.no

Abstract. According to the literature, normal ageing is associated with a decline in sensory, perceptual, motor and cognitive abilities. When designing applications for elderly people, it is crucial to take into consideration the decline in functions. For this purpose, gesture-based applications that allow for direct manipulations can be useful, as they provide natural and intuitive interactions. This paper examines gesture-based applications for the elderly and studies that have investigated these applications, and it identifies opportunities and challenges in designing such applications.

Keywords: Gesture, elderly, direct manipulation, accessibility.

1 Introduction

Many industrialised countries are experiencing a huge demographic change, where the proportion of the elderly population is increasing to unprecedented levels, and this population will continue to grow significantly in the future. It is widely accepted that more research is needed to address this issue. With information technologies becoming commonplace in society, the opportunity and necessity for elderly people to access these technologies in their everyday activities have been increasing. Human-computer interaction must be designed and implemented so that age-related challenges in functional ability, such as perceptual, cognitive and motor functions, are taken into account.

In the last decade, much attention has been devoted to understanding and accommodating the needs of the elderly with respect to interaction with computers through a keyboard and mouse. Recent years have seen the increasing popularity of gesture-based applications, where users use the movements of the hands, fingers, head, face and other parts of the body to interact with virtual objects. Furthermore, studies have been carried out to investigate how older users use gesture inputs in their interactions with information technologies. Compared with mouse and keyboard inputs, gesture interfaces have the advantage of simplicity; they require less learning time. For older users, who may operate a mouse or keyboard with limited speed and accuracy, the gesture interfaces can be attractive and make applications more accessible. In this paper, we survey and characterise existing research on gesture-based applications for the elderly and identify the challenges and discuss the research opportunities that those challenges offer.

2 Gesture Interfaces

The input methods used for human-computer interaction are important because the usability of an input method affects the overall effectiveness of an interactive system. A gesture is non-verbal communication made with the hand, finger, head, face or other part of the body. Gestures, independently or in combination with verbal communication, are used commonly to communicate messages. With gestures, humans can directly interact with machines without the extra layer of mechanical devices, such as a mouse or keyboard. This type of interaction is considered more natural and intuitive because humans learn to use gestures from childhood.

Earlier work in gesture interfaces focused on the use of gestures for editing purposes[1]. These interfaces usually involved the user writing directly on the surface of a display with a stylus. The further development of gesture interfaces includes using gloves with sensors to identify hand and finger movements [2], using special suits with sensors to track full body movements [3], finger gestures on a single touch screen and multi-touch tabletops (e.g. Microsoft Surface, Apple touchpad), using an accelerometer to track movements (e.g. Wii remote, [4, 5]) as well as face recognition and motion tracking with no sensors on the body or controller in the hands (e.g. Sony EyeToy, Microsoft Kinect, Flutter and GestureTek).

Summarising the different types of gesture interfaces and technical approaches, Karam and Schraefel [6] proposed a taxonomy of gestures in human-computer interaction. They categorised the gestures in terms of four key elements: gesture styles, the application domains to which they are applied, input technologies and output technologies used for implementation. Bhuiyan and Picking [7] reviewed the history of gesture controlled user interfaces including types of gestures, their users, applications, technology and the issues addressed over the past 30 years, and they identified trends in technology, application and usability. In their paper, Bhuiyan and Picking [7] also provided a research background for gesture interfaces for elderly or disabled people.

3 Characteristics of Elderly People

Age is a surrogate variable that only loosely predicts the amount of disablement of any particular older person [8]. There are many older people who can and do use applications designed for younger users, but it is clear that learning and using full-featured applications designed for younger people is very difficult for many older people. The literature has shown that normal ageing is associated with a decline in sensory, perceptual, motor and cognitive abilities. The older users that this paper focuses on are those who suffer from these negative effects of ageing and experience declines in different abilities.

Physical Characteristics. The effects of ageing on motor abilities generally include slower response times, coordination reduction and a loss of flexibility [9]. A decline in motor abilities, especially fine motor skills, is a problem for many older people

when using mobile phones or laptop computers with integrated mice. Using a computer mouse can be difficult for older users because it requires good hand-eye coordination [10]. Some older users find the double-click very difficult, if not impossible. Reduced motor skills also cause more errors during fine movements, especially when other cognitive functions are required at the same time [11]. Elderly people often confuse the right-click with the left-click while they are at the same time trying to attend to the computer screen.

Visual perception worsens with ageing. The size of the visual field decreases, which leads to a loss of, for example, peripheral vision, colour vision, contrast detection and dark adaptation [12]. In addition, hearing ability declines to 75% for people between 75 and 79 years of age [12, 13]. Elderly people also become more easily distracted by details or noises. They have difficulty maintaining attention on more than one aspect at once [14]. Ageing also causes the short-term memory to retain fewer items, the working memory to be less efficient and the perspective memory, that is, the ability to remember, to be reduced when complex tasks are involved [15]. The combination of reduced vision, hearing, memory and mobility contributes to a loss of confidence, which may cause isolation and depression and lead to difficulty in learning, and sometimes hinder the use of new technologies.

Emotion Aspects and Social Engagement. It is often claimed that elderly people are reluctant to make use of state-of-the-art technology, even if they have enough cognitive and physical abilities to do so. Technologies tend to make them uncomfortable. They do not seem to trust their own capabilities and are often afraid of making mistakes that they think may cause damage to the system. This computer anxiety can be reduced after one is taught computer-based skills [16] and when the applications have a higher learnability and a higher recognisability.

Elderly people often experience social isolation. Studies have shown that social disconnectedness (e.g. small social network, infrequent participation in social activities) and perceived isolation (e.g. loneliness, perceived lack of social support) have distinct associations with physical and mental health among older adults [17]. Hence, social activities and engagement are found to be important for the well-being of elderly people.

4 Research on Gesture-Based Applications for the Elderly

Gesture interfaces may make applications more attractive and friendly to older users because they are natural and intuitive, they require minimal learning time and they lead to a high degree of user satisfaction. The touch screen has been suggested as a suitable input device for elderly users because it is easy to learn and operate [18, 19]. There is a large body of work on single finger touch screen applications for older users, but relatively little work has been done on the use of multi-touch, hand, face and body gestures for interacting with applications. In this section, we categories existing research on gesture interface applications for elderly people based on their purposes, including training and rehabilitation, entertainment and social activities, and independent living.

Training and Rehabilitation. Gesture-based applications can be used for the training and rehabilitation of specific body parts. In such applications, gesture interfaces are usually combined with simulation or virtual reality [20]. The user's gestures are translated into the movements of an avatar in the virtual world. For example, when an older user moves his/her arms, he/she can see in the virtual world the avatar doing the same movements in real-time. Such interfaces are especially useful for elderly users who need to rehabilitate specific body parts. They can guide the users through clinician-prescribed interactive rehabilitation exercises, games and activities that can target these body parts.

A number of multi-touch tabletop applications targeting elderly users have been developed for training purposes. In the HERMES ('Cognitive care and guidance for active aging') project, several cognitive training games were implemented on multi-touch tables [21]. Apted et al. [22] and Al Mahmud et al. [23] proposed a list of design guidelines for tabletop-based applications for the elderly. Following these guidelines, Annett et al. [24] developed a suite of five motor-based rehabilitation activities for older users. For an overview of multi-touch tabletop applications for training and rehabilitation for the elderly, see [25].

Entertainment and Social Activities. Due to the decline in abilities that comes with ageing, elderly people tend to live an isolated existence. Therefore, the social isolation of the elderly is becoming a pressing problem. To improve their quality of life, it is essential that elderly people have an active social and physical life, which is often called 'active ageing'[26]. An increasing number of multi-touch tabletop applications promote entertainment and social interactions among elderly people. Hollinworth and Hwang [27] designed an e-mail application on a multi-touch table for elderly users, allowing the user to use finger gestures to manipulate objects on the table. 'Familiarity' as a design principle was adopted in the design of the interface, where tools were provided in the form of familiar visual objects and manipulated by finger gestures, just like their real-world counterparts, rather than with buttons, icons and menus. The formative evaluation showed that three of the four participants were able to carry out some of the basic e-mail tasks with no prior training and little or no help. Sharetouch [28] is another multi-touch tabletop application designed to enable social interaction among the elderly living in a community. For an overview of multi-touch tabletop applications for social interactions among the elderly, see [25].

Several Nintendo Wii games, especially sport games such as Bowling, have also been used in residential homes [29-31] to promote social and physical activities amongst elderly people. Some research has shown the positive effects of Wii games on elderly people's physical health [32] and mental well-being [32, 33]. Volda and Greenberg [30] reported the results of their qualitative study on Wii games serving as a meeting place for diverse people. Their study participants were the residents of a retirement community, who played the Wii Bowling game. An interesting finding from this study was that a more experienced elderly player advised the other players to bowl with the Wii remote held upside down, so that they would only see the one relevant button. The Wii remote has many buttons which are not needed to play this game, but they interrupt the normal sequence by opening up menus when accidentally

pressed. This was unexpected for the older players, and when these menus would pop up, they were afraid that they had destroyed the game. To prevent the elderly players from accidentally pressing the irrelevant buttons, Neufeldt [29] covered these buttons in his study.

Neufeldt [29] conducted his study with members of a fitness programme in a retirement home. Six participants played four rounds of Wii Bowling with each round associated with a regular fitness session that occurred once a month. Observation was the main data collection method. In addition to the irrelevant button on the Wii remote, this study found that the game required a high level of attention from the elderly players, because the button must be released at the right moment, and this requires good hand-eye coordination. The explanations and help from the researchers, the hints from other elderly players and the sounds of the game caused considerable stress for the participants, which resulted in less fun and more reluctance to play. The researcher, however, observed development in the capabilities of the elderly participants from round to round, which indicated that Wii games could help improve coordination capabilities and encourage the elderly to move their arms.

Harley et al. [31] conducted a longitudinal study of older people's use of the Wii in Shelter Housing over a period of one year. Data were collected using observations, interviews and video recording. Through interaction analysis, the study highlighted how older people actively constructed the sense of a meeting place by gaining control over the space and engaging in the social processes. Harley et al. [31] concluded their paper by presenting five design implications and guidelines for encouraging appropriation and empowerment among older people through game play in communal housing settings.

Jung et al. [32] conducted a six-week comparative study to examine the impact of playing Wii games on the psychological and physical well-being of the elderly in a long-term care facility. The experiment group included 45 residents who played Wii games; the control group played traditional board games. The results showed that playing Wii games had a positive impact on the overall well-being of the elderly compared to the control group.

Gerling et al. [34] conducted two studies using Microsoft Kinect. The first study focused on the suitability of the gesture set for institutionalised older people. The gesture set included four static body gestures and four dynamic body gestures. Seventeen elderly people from 60 to 90 years of age participated in the study. The second study focused on testing a game using body gestures. Twelve elderly people from 60 to 91 years of age participated in the study. Both qualitative data (questionnaire and observation) and quantitative data (performance metrics) were collected. Gerling et al. [34] concluded the paper by presenting seven guidelines for full-body interactions in games for elderly people.

Independent Living. It is commonly recognised that elderly people can benefit from the use of information and communication technologies in their homes to allow for longer independent living. However, in order to enable the independent living of elderly people, many challenges need to be addressed [35]. Various assistive technologies and services have been developed to support independent living in

different aspects of life, including safety, health and wellness and social connectedness [36]. However, few systems have taken advantage of gesture technology. As a part of IBM's accessibilityWorks project, a gesture interface called TouchFree Switch [37] was developed for older users, allowing the user to interact with the Mozilla web browser by the tip of the head, a shrug of the shoulder, a finger movement or any other body movement. Furthermore, the TouchFree Switch interface allows users to choose their own gestures and associate them with tasks. Jia et al. [38] developed a hands-free intelligent wheelchair control with head gestures for the elderly and people with disabilities. The recognised head gestures are used to generate motion control commands to the motion controller in the wheelchair so that it can control the motion of the wheelchair according to the user's intentions. The preliminary results showed that the gesture interface was very useful for the users who have restricted limb movements.

The Gesture Pendant [39] is a wearable device allowing older users to control home automation systems via hand gestures. Thus, home devices to control, for example, entertainment equipment and room lighting can be controlled by hand movements. Eight standard hand gestures ('horizontal pointed finger up', 'horizontal pointed finger down', 'vertical pointed finger left', 'vertical pointed finger right', 'horizontal flat hand down', 'horizontal flat hand up', 'open palm hand up' and 'open palm hand down') were defined as control gestures. In addition, the system allowed the users to self-define gestures for the tasks, for example, 'fire on', 'fire off', 'door open', 'door close', 'window up', and 'window down'.

Open Gesture [40] allows older people to use hand gestures to perform a diverse range of tasks at home via a television interface. After running the application (which could be initiated by selecting a pre-configured television channel), the user could see his/her image on the television screen, which was filmed through a connected webcam. The user could point at different objects using hand gestures to perform various tasks, such as making a telephone call, playing 'brain training' games, controlling the computer or home environment and social networking.

5 Challenges and Opportunities

Gesture interfaces provide realistic and affordable opportunities and offer some potential for improving the independence and quality of life of elderly people. However, there remain significant challenges to overcome.

Technological Challenges. Due to the anxiety that elderly users experience when interacting with technologies, gesture-based applications must provide reliable and easily performable gestures. Gesture recognition is a challenging task and essential for gesture interfaces. Techniques such as hidden Markov models (HMMs), particle filtering and condensation, finite-state machine (FSM) and artificial neural networks (ANNs) have been adopted for recognising hand and arm gestures [41, 42]. Moreover,

new and improved tools and techniques must be at the centre of research in order to increase the reliability and accuracy of gesture recognition systems. Many gesture-based applications have been developed for user groups besides the elderly and have proved useful to these user groups. For example, Stomp [43] is an application designed to support social and physical interactions for people with intellectual disability. The mini-games in this application can be easily adapted for elderly people in communities or residential homes. A gesture-based application should be able to lower the resistance of the users and offer a clear benefit, whether physical, medical or emotional, in order for elderly users to accept it. This must be achieved on several levels. On the design level, the application must be adapted to older users' physical and perceptual characteristics. The interface should offer a certain degree of familiarity to overcome reservations. On the function level, the benefit of using the application must be appreciable in order to provide a motivation for its use. A balance must be established between intuitive use and practical learning.

Methodological Challenges. The results from earlier studies have shown that there is a wide gap between the young designers' personal experience and the experiences of the older users. In order to design and implement useful and accessible applications for elderly people, it is important to increase the awareness of the characteristics of elderly people among the designers and developers of the applications. Thus, a user-centred approach should be adopted in the design and development process. Designers cannot only follow guidelines; they must also involve older users from the early stage of the development process. Due to the gap in experience, those designing for older users will remain dependent on testing with a range of older users in order to verify the assumptions made in their designs. However, older people provide far greater challenges to user-centred design than more traditional user groups [44]. Newell [44] proposed different methodologies for involving older adults in the design process, including the use of theatre. The studies on elderly people with gesture-based applications have focused mainly on attitudes and subjective evaluation. Owing to the inherent limitation of subjective measurements, it is important to use objective measures, such as performance data. Although previous research has suggested that gesture interfaces may be especially easy for older users to use, as they allow for direct manipulation, until now, few researchers have systematically investigated the usability of gesture-based applications for older users. Hence, systematic studies are necessary to understand how elderly people use these applications and to establish design recommendations for such applications.

In order to confirm whether gesture interfaces are acceptable for actual use, long-term investigation following the adoption of the technology is important. Longitudinal studies should focus on the usability of gesture interfaces, the acceptance of gesture interfaces by elderly users including attributes and motivations as well as their performance and improvement in sensory, perceptual, motor and cognitive abilities.

6 Conclusion

Information and communication technologies (ICT) have been proposed as useful for offsetting the negative effects of physical, cognitive and social ageing. However, the uptake of ICT by elder people is rather low. Innovative interaction technologies, such as gesture technology, have great potential for improving the accessibility of interactive systems to elderly users. Despite the limited research, the evidence suggests that gesture technology is an applicable and practical technology for this user group. However, we still need to understand how to take advantage of this technology to provide the best possible support for elderly people. Future projects could pursue enquiry in many directions in order to fully explore the potential of gesture technology.

References

1. Rhyne, J.: Dialogue Management for Gestural Interfaces. *ACM SIGGRAPH Computer Graphics* 21, 137–142 (1987)
2. Zimmerman, T.G., Lanier, J., Blanchard, C., Bryson, S., Harvill, Y.: A hand gesture interface device. *SIGCHI Bull.* 18, 189–192 (1986)
3. Fitzgerald, D., Foody, J., Kelly, D., Ward, T., Markham, C., McDonald, J., Caulfield, B.: Development of a wearable motion capture suit and virtual reality biofeedback system for the instruction and analysis of sports rehabilitation exercises. In: *Proc. of IEEE Conf. Eng. Med. Biol. Soc.*, pp. 4870–4874. IEEE (2007)
4. Sawada, H., Hashimoto, S.: Gesture Recognition Using an Accelerometer Sensor and its Application to Musical Performance Control. *Electronics and Communications in Japan Part 3*, 9–17 (2000)
5. Amini, N., Sarrafzadeh, M., Vahdatpour, A., Xu, W.: Accelerometer-based on-body sensor localization for health and medical monitoring applications. *Pervasive Mob. Comput.* 7, 746–760 (2011)
6. Karam, M., Schraefel, M.C.: A taxonomy of gestures in human computer interactions. Technical report, Electronics and Computer Science, University of Southampton (2005)
7. Bhuiyan, M., Picking, R.: Gesture Control User Interface, what have we done and what's next? In: Bleimann, G.U., Doland, S.P., Furnell, M.S., Grout, V. (eds.) *Proceedings of the Fifth Collaborative Research Symposium on Security, E-learning, Internet and Networking (SEIN-2009)*, pp. 59–68. University of Plymouth, Darmstadt (2009)
8. Hawthorn, D.: How universal is good design for older users? *ACM SIGCAPH Computers and the Physically Handicapped*, 38–45 (2002)
9. Rogers, W.A., Fisk, A.D., Mead, S.E., Walker, N., Cabrera, E.F.: Training older adults to use automatic teller machines. *Hum. Factors* 38, 425–433 (1996)
10. Walker, N., Philbin, D.A., Fisk, A.D.: Age-related differences in movement control: adjusting submovement structure to optimize performance. *J. Gerontol. B Psychol. Sci. Soc. Sci.* 52, 40–52 (1997)
11. Charness, N., Bosman, E.A.: Human factors and design for older adults. In: Birren, J.E., Schaie, K.W. (eds.) *Handbook of the Psychology of Aging*, pp. 446–463. Academic Press, New York (1990)
12. Fozard, J.L.: Vision and hearing in aging. In: Birren, J.E., Schaie, K.W. (eds.) *Handbook of the Psychology of Aging*, pp. 150–170. Academic Press, New York (1990)

13. Schieber, F.: Aging and the senses. In: Birren, J.E., Sloan, R., Cohen, G. (eds.) *Handbook of Mental Health and Aging*, pp. 251–306. Academic Press, New York (1992)
14. McDowd, J.M., Craik, F.I.: Effects of aging and task difficulty on divided attention performance. *J. Exp. Psychol. Hum. Percept. Perform.* 14, 267–280 (1988)
15. Craik, F.I.M., Jennings, J.J.: Human memory. In: Craik, F.I.M., Salthouse, T.A. (eds.) *Handbook of Aging and Cognition*, pp. 51–110. Erlbaum, Hillsdale (1992)
16. Ellis, D., Allaire, J.C.: Modeling computer interest in older adults: the role of age, education, computer knowledge, and computer anxiety. *Hum. Factors* 41, 345–355 (1999)
17. Cornwell, E.Y., Waite, L.J.: Social Disconnectedness, Perceived Isolation, and Health among Older Adults. *J. Health Soc. Behav.* 50, 31–48 (2009)
18. Yarnold, P.R., Stewart, M.J., Stille, F.C., Martin, G.J.: Assessing functional status of elderly adults via microcomputer. *Percept. Mot. Skills* 8, 689–690 (1996)
19. Tobias, C.L.: Computers and the Elderly: A Review of the Literature and Directions for Future Research. In: *Proc. of the Human Factors Society 31st Annual Meeting*, pp. 866–870. Human Factors Society, Santa Monica (1987)
20. Kallio, S., Kela, J., Mäntyjärvi, J., Plomp, J.: Visualization of Hand Gestures for Pervasive Computing Environments. In: *Proceedings of the Working Conference on Advanced Visual Interfaces (AVI 2006)*, pp. 480–483. ACM, New York (2006)
21. Facal, D., Buiza, C., González, M.F., Soldatos, J., Petsatodis, T., Talantzis, F., Urdaneta, E., Martínez, V., Yanguas, J.J.: Cognitive Games for Healthy Elderly People in a Multitouch Screen. In: *Proc. of the International IDRT4ALL, Barcelona, Spain (2009)*
22. Apted, T., Kay, J., Quigley, A.: Tabletop Sharing of Digital Photographs for the Elderly. In: *Proc. of CHI 2006*, pp. 781–790 (2006)
23. Al Mahmud, A., Mubin, O., Shahid, S., Martens, J.: Designing and Evaluating the Tabletop Game Experience for Senior Citizens. In: *Proc. of NordiCHI 2008*, pp. 403–406 (2008)
24. Annett, M., Anderson, F., Goertzen, D., Halton, J., Ranson, Q., Bischof, W.F., Boulanger, P.: Using a multi-touch tabletop for upper extremity motor rehabilitation. In: *Proc. of the 21st Annual Conference of the Australian Computer-Human Interaction Special Interest Group: Design*, pp. 261–264. ACM, New York (2009)
25. Loureiro, B., Rodrigues, R.: Multi-Touch as Natural User Interface for Elders: A survey. In: *Proc. of 6th Iberian Conference on Information Systems and Technologies CISTI 2011, Chaves, Portugal*, pp. 573–578 (2011)
26. OECD: *Reforms for an Ageing Society*. OECD Publishing, Paris (2000)
27. Hollinworth, N., Hwang, F.: Investigating familiar interactions to help older adults learn computer applications more easily. In: *Proceedings of the 25th BCS Conference on Human-Computer Interaction (BCS-HCI 2011)*, pp. 473–478. British Computer Society, Swinton (2011)
28. Tsai, T.-H., Chang, H.-T., Chang, Y.-M., Huang, G.-S.: Sharetouch: A system to enrich social network experiences for the elderly. *J. Syst. Softw.* 85, 1363–1369 (2012)
29. Neufeldt, C.: Wii play with elderly people. *International Reports on Socio-Informatics (IRSI)* 6 (2009)
30. Voida, A., Greenberg, S.: Wii all play: the console game as a computational meeting place. In: *Proc. of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2009)*, pp. 1559–1568. ACM, New York (2009)
31. Harley, D., Fitzpatrick, G., Axelrod, L., White, G., McAllister, G.: Making the Wii at home: game play by older people in sheltered housing. In: Leitner, G., Hitz, M., Holzinger, A. (eds.) *USAB 2010. LNCS*, vol. 6389, pp. 156–176. Springer, Heidelberg (2010)

32. Jung, Y., Li, K.J., Janissa, N.S., Gladys, W.L.C., Lee, K.M.: Games for a better life: effects of playing Wii games on the well-being of seniors in a long-term care facility. In: Proc. of the Sixth Australasian Conference on Interactive Entertainment (IE 2009). ACM, New York (2009)
33. Theng, Y.-L., Chua, P.H., Pham, T.P.: Wii as entertainment and socialisation aids for mental and social health of the elderly. In: Proc. CHI 2012 Extended Abstracts on Human Factors in Computing Systems (CHI EA 2012), pp. 691–702. ACM, New York (2012)
34. Gerling, K., Livingston, I., Nacke, L., Mandryk, R.: Full-body motion-based game interaction for older adults. In: Proc. of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2012), pp. 1873–1882. ACM, New York (2012)
35. Edwards, W.K., Grinter, R.E.: At Home with Ubiquitous Computing: Seven Challenges. In: Abowd, G.D., Brumitt, B., Shafer, S. (eds.) *UbiComp 2001*. LNCS, vol. 2201, pp. 256–272. Springer, Heidelberg (2001)
36. Alwan, M., Wiley, D.: *State of Technology in Aging Services*. Report, Center for Aging Services Technologies (2007)
37. Hanson, V.L., Brezin, J.P., Crayne, S., Keates, S., Kjeldsen, R., Richards, J.T., Swart, C., Trewin, S.: Improving web accessibility through an enhanced open-source browser. *IBM Systems Journal* 44, 573–588 (2005)
38. Jia, P., Hu, H., Lu, T., Yuan, K.: Head gesture recognition for hands-free control of an intelligent wheelchair. *Journal of Industrial Robot* 34, 60–68 (2007)
39. Gandy, M., Starner, T., Auxier, J., Ashbrook, D.: The Gesture Pendant: A Self-illuminating, Wearable, Infrared Computer Vision System for Home Automation Control and Medical Monitoring. In: *Proceedings of the 4th IEEE International Symposium on Wearable Computers (ISWC 2000)*, pp. 87–94. IEEE Computer Society, Washington, DC (2000)
40. Bhuiyan, M., Picking, R.: A Gesture Controlled User Interface for Inclusive Design and Evaluative Study of Its Usability. *Journal of Software Engineering and Applications* 4, 513–521 (2011)
41. Mitra, S., Acharya, T.: Gesture Recognition: A Survey. *IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews* 37, 311–324 (2007)
42. Khan, R.Z., Ibraheem, N.A.: Survey on Gesture Recognition for Hand Image Postures. *International Journal of Computer and Information Science* 5, 110–121 (2012)
43. Wyeth, P., Summerville, J., Adkins, B.: Stomp: an interactive platform for people with intellectual disabilities. In: Romão, T., Correia, N., Inami, M., Kato, H., Prada, R., Terada, T., Dias, E., Chambel, T. (eds.) *Proceedings of the 8th International Conference on Advances in Computer Entertainment Technology (ACE 2011)*. ACM, New York (2011)
44. Newell, A.F.: User sensitive design for older and disabled people. In: Helal, S., Mokhtari, M., Abdularazak, B. (eds.) *The Engineering Handbook of Smart Technology for Ageing, Disability and Independence: Computer and Engineering for Design and Applications*, pp. 787–892. Wiley, New Jersey (2008)

MOBAJES: Multi-user Gesture Interaction System with Wearable Mobile Device

Enkhbat Davaasuren and Jiro Tanaka

1-1-1 Tennodai, Tsukuba, Ibaraki 305-8577 Japan
{enkhee, jiro}@iplab.cs.tsukuba.ac.jp

Abstract. When people collaborate with multiple large screens, gesture interactions will be used widely. However, in conventional methods of gesture interaction, when there are multiple users, simultaneous interaction is difficult. In this study we have proposed a method using a wearable mobile device which enables multi-user and hand gestures only interactions. In our system, the user wears a camera-equipped mobile device like a pendant, and interacts with a large screen.

Keywords: Gesture, Gestural Interface, Large Screen, Mobile, Wearable Device, Multi-User.

1 Introduction

In the past years, large screen has been used more and more in various locations and situations, and their use will likely increase in the future. Many researchers have been performing research about large screen interaction methods. One of the most used interaction methods is Gesture Interaction, which is a method where the user can use body or hand gestures to interact with large screen. There are many types of gesture interaction systems, and each of them has good and weak points in multi-user interactions. In this research, we consider the hand gesture methods, and propose a hybrid interaction system that can work in a more stable manner in multi-user interactions (Fig. 1).



Fig. 1. System image

1.1 Gesture Interface

Two kinds of gesture interfaces exist: wearable and non-wearable. In wearable interfaces, gestures are recognized by a sensor which is installed on the user's hand or body. In non-wearable interfaces, gestures are recognized by a camera which is attached to the large screen. The advantages and disadvantages of these two types are quite opposite. The advantage of non-wearable interfaces is that users do not need to wear any devices or markers, and this makes the system more mobile and easy to use. On the other hand, in a multi-user interaction, non-wearable interfaces are not so suitable. For example, problems such as calculation cost of gesture recognition or difficulty of identifying different users may occur. However, wearable interfaces also have disadvantages, like users needing to wear or hold in hand devices or markers. They are more suitable to multi-user interactions. This is because using a gesture recognizing device for each user makes it easier to identify the user, and the number of users will not affect the system calculation cost.

1.2 Proposal System

Our proposed system is a hybrid system, which applied both wearable and non-wearable advantages. We approached by wearing camera-equipped mobile device like a pendant as a gesture recognition device. In our system, user makes hand gestures in front of device which worn like a pendant, and make gesture interactions while seeing cursors of gestures on the large screen (Fig. 2). Since user does not need to wear device in hand, user can use both two hands freely, and can move freely even during the interaction. Also, user can concentrate to the interactions without thinking about device. Because each user has own device, user number will not affect to the system calculation cost.

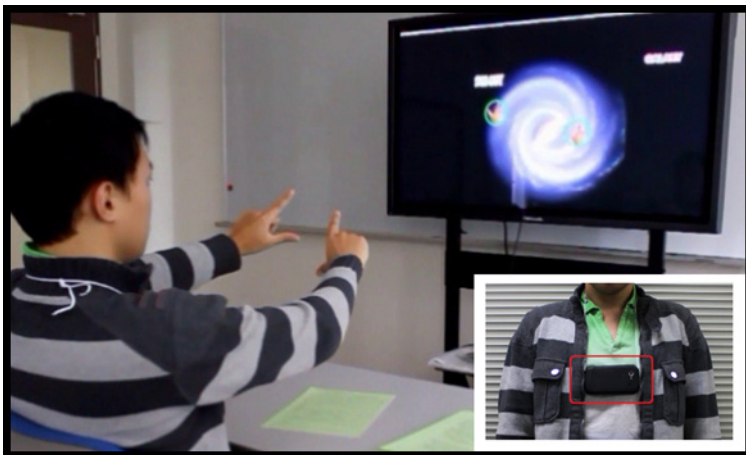


Fig. 2. MOBAJES system

2 Related Work

In Gesture Pendant [1], researchers proposed new approach to detect hand gestures, but they did not consider multi-user. One important difference is that our system provides GUI feedback by using large screen during gesture interaction, while enabling more rich interactions to the user.

There are also many non-wearable gesture interaction systems such as [2], [3], [4], [5] and [6]; however, they still present multi-user interaction issues. In these systems, the whole recognition process is calculated in one place, and that makes the system unstable in case of multi-user interactions. We applied wearable approach to address these issues.

Systems about gesture interactions for public large screens ([7], [8], [9] and [10]) have also been developed. However, in these systems, the user needs to hold a mobile device in hand when interacting with large screen, which poses a burden to the user. In our system, user wears mobile device like a pendant, and thus does not burden the user.

The most related work to our system is Sixth Sense [11]. In the Sixth Sense system, user can display information on the other objects such as walls, by using a wearable projector, and make gestural interactions using markers of hand. In our system, the user can obtain more clear and rich information through the large screen, and can interact with bare hand gestures without markers.

3 MOBAJES Interactions

We developed a prototype system as a simple image manipulation system. In our prototype, user can manipulate the image files on the large screen using hand gestures, by wearing a mobile device like a pendant (Fig. 2). During the interaction, the user can see a cursor (Fig. 3 b) on the large screen as a feedback of gesture (Fig. 3 a). This will help users to understand each other's intention during multi-user interaction.

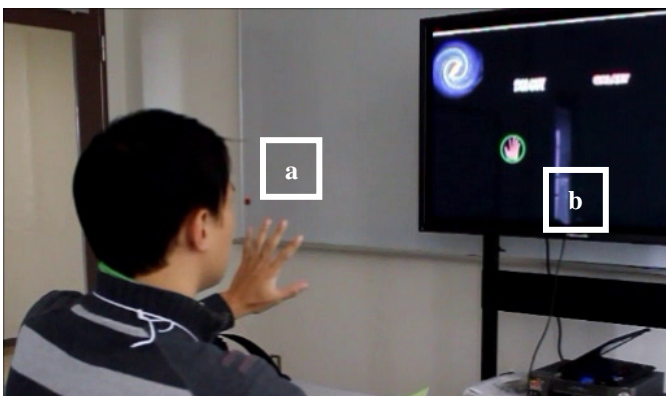


Fig. 3. Gesture feedback (a: user gesture, b: cursor on the large screen)

In this system, user can use 4 kinds of gesture such as “grab”, “release”, “point” and “L-letter” (Fig. 4). Those of the same shaped cursors will appear on the large screen.

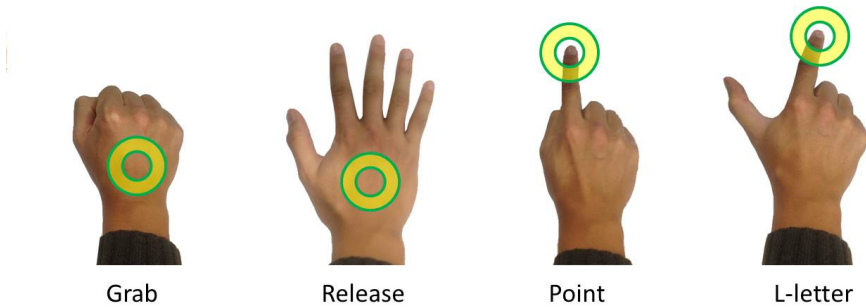


Fig. 4. Gesture types (Round markers represent the targeting point of each gestures)

Using these gestures, we implemented several basic interactions for our prototype, such as drag & drop, zoom & rotate and file share.

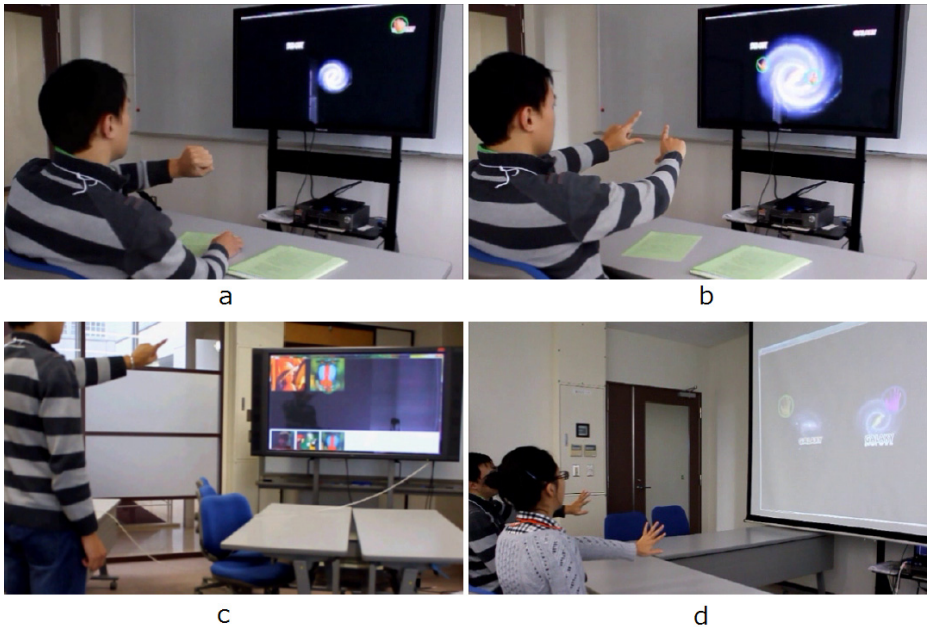


Fig. 5. Interactions (a: Drag & drop, b: Zoom & Rotate, c: File share, d: Multi-user interaction)

Drag & Drop. After user hovers over the target image file (using the cursor), he/she can drag the file with the “Grab” gesture, and can drop it with the “Release” gesture (Fig. 5-a).

Zoom & Rotate. User can zoom and rotate the file by using two hands (Fig. 5-b). To perform it, user needs to hover over the target image with tow hand’s “Point” gesture, and while keeping that position, user needs to change the gesture to “L-letter”. After that, user can zoom and rotate the file by changing the distance and the direction of two hands.

File Share. User can select the file on the screen using one-handed “Point” and “L-letter” gestures to copy it to his/her mobile device. To perform it, user need to hover over the target image file on the screen with “Point” gesture, and change the gesture to “L-letter”. In reverse, user can display thumbnails of the image files of his/her mobile device on the large screen (Fig. 5-c), by changing gesture from “Release” to “L-letter”. After displaying the thumbnails on the screen, user can copy and put the original file to large screen from mobile device by same gesture.

Multi-User Interaction. All these interactions can be performed in multi-user interactions as well (Fig. 5-d). Users can interact separately and collaboratively with each other, while knowing each other’s intention.

4 Implementation

We have implemented our system using a camera-equipped Android mobile device, a large screen and Wi-Fi environment (Fig. 6).

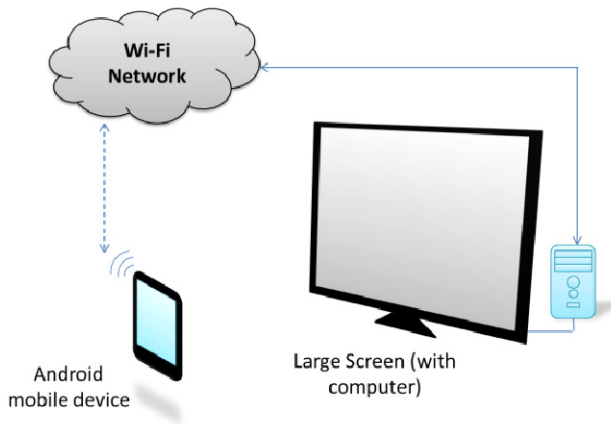


Fig. 6. System structure

4.1 Communications

Mobile devices and large screen will connect to one server application on the network, and communicate with each other using socket connection. The gesture information needs to be transferred in real time, so it is transferred by UDP protocol,

between the mobile device and the large screen. Other information such as commands and files are transferred using the TCP protocol.

4.2 Gesture Recognition

The whole gesture recognition process is calculated by the mobile device. We used OpenCV¹ library and skin-color based method for recognizing gesture information. In order to recognize a hand gesture, we first detect skin-color areas (Fig. 7-b) from camera capture (Fig. 7-a). Using noise removal method, we can obtain clearer skin-color regions (Fig. 7-c).

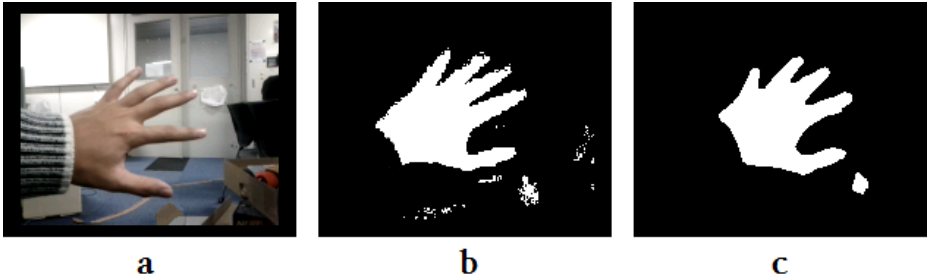


Fig. 7. Detecting skin color region (a: camera capture image, b: skin-color region with noise, c: clear skin-color region)

After that, we extract the contours of each skin-color regions (Fig. 8-a), and filter them by the area to obtain hand region contour (Fig. 8-b). Next, we extract the convex hull of hand region to detect finger-like parts. As can be noticed in the figure, finger tips belong both to the contour and the convex hull of the hand region (red parts in Fig. 8-c). We can use this fact to decrease the calculation cost and make recognition faster.

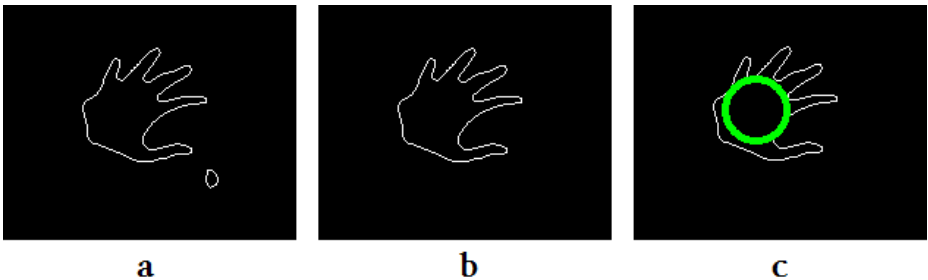


Fig. 8. Detection of hand region (a: contours of skin-color regions, b: contour of hand region, c: contour and convex hull of hand region)

¹ <http://opencv.org>

Next, we detect finger-like parts by using the angle of every tree points on the contour (Fig. 9-a). If the angle $\angle\theta$ is less than 30 degrees, it indicates the point P_i is the potential point of fingertip. Finally, we calculate the center of potential fingertip points as a real fingertip point (Fig. 9-b).

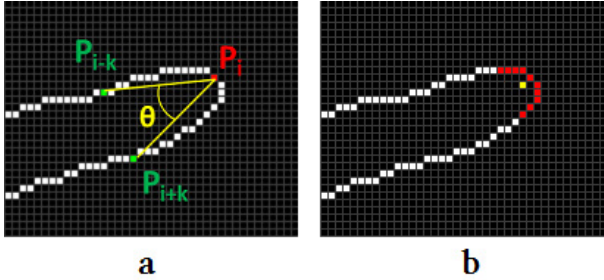


Fig. 9. Detection of fingertip (a: the angle between specific 3 points, b: found fingertip)

4.3 Noise Removal

To remove noise in gesture recognition process, we used the fact that the distance between user's hand and camera is almost constant (Fig. 10). If the distance is almost constant, we can assume the area size of hand region must be constant. By filtering the area size of all found regions, noises such as small size regions will be removed.

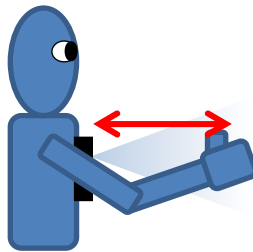


Fig. 10. Distance between user's hand and camera

We also used the natural fact that the hand region cannot be long and narrow shaped. To filter and remove long and narrow regions, we used the distance between the center point and the point closest to the center of each region. If the distance is less than specific threshold it means this shape is long and narrow and has to be removed.

After filtering the regions found by these limitations, we can obtain more clear hand regions for the further process.

5 Preliminary Evaluation

We performed a preliminary experiment to evaluate our system in multi-user interactions. In order to evaluate our system we asked two users to complete the given task in case of both single-user and multi-user interactions. After the experiment, we asked them about the difference between single-user and multi-user interactions. We also measured the amount of time needed to complete the task.

The task consisted of simply dragging and dropping a given picture to specified positions. To compare fairly, we used only one hand gesture for this task.

Our result shows that the performance of single-user interaction is almost same in both systems. And users said there is no difference between single-user interaction and multi-user interaction. Furthermore, in a multi-user interaction, knowing other user's intention by the cursor, it was easy to collaborate and avoid the collision.

In our next work, we will perform user study to know the error rates in each interactions. And to prove usefulness of our system, we will compare our system with a non-wearable gesture interface for large screen.

6 Summary and Future Work

In this study we proposed MOBAJES, a system which can intuitively interact with a large screen using a mobile device, and we implemented a prototype system. By wearing camera-equipped mobile device like a pendant, and performing hand gestures in front of the camera, the user can interact with large screen by gesture interaction. Since each user has own gesture recognition device, the gesture recognition cost does not affect the whole system cost, and user identification becomes very easy.

But, since we use skin-color detection method to recognize hand gestures, recognition accuracy easily affected by lighting of environment. We believe we need a more robust recognition algorithm suited to dynamic lighting change. In our future work, we will improve the recognition accuracy by implementing a more dynamic algorithm. We also intend to try a different device such as a depth sensor. Furthermore, we found that the gestures we use in our system can be tiring in case of a task taking a longer time. To address this problem, we need to consider easier gestures which users can perform easily and naturally without stress.

References

1. Gandy, M., Starner, T., Auxier, J., Ashbrook, D.: The Gesture Pendant: A Self-illuminating, Wearable. In: *Infrared Computer Vision System for Home Automation Control and Medical Monitoring*, ISWC 2000, pp. 87–94. IEEE Computer Society (2000)
2. Boulabiar, M.-I., Burger, T., Poirier, F., Coppin, G.: A low-cost natural user interaction based on a camera hand-gestures recognizer. In: Jacko, J.A. (ed.) *Human-Computer Interaction, Part II*, HCII 2011. LNCS, vol. 6762, pp. 214–221. Springer, Heidelberg (2011)
3. Shi, J., Zhang, M., Pan, Z.: A real-time bimanual 3D interaction method based on bare-hand tracking. In: *MM 2011*, pp. 1073–1076. ACM (2011)

4. Bragdon, A., DeLine, R., Hinckley, K., Morris, M.R.: Code space: touch + air gesture hybrid interactions for supporting developer meetings. In: ITS 2011, pp. 212–221. ACM (2011)
5. Argyros, A.A., Lourakis, M.I.A.: Vision-based interpretation of Hand Gestures for Remote Control of a Computer Mouse. In: Huang, T.S., Sebe, N., Lew, M., Pavlović, V., Kölsch, M., Galata, A., Kisačanin, B. (eds.) HCI/ECCV 2006. LNCS, vol. 3979, pp. 40–51. Springer, Heidelberg (2006)
6. Clark, A., Dnser, A., Billinghamurst, M., Piumsomboon, T., Altimira, D.: Seamless interaction in space. In: Proceedings of the 23rd Australian Computer-Human Interaction Conference (OzCHI 2011), pp. 88–97. ACM (2011)
7. Ballagas, R., Rohs, M., Sheridan, J.G.: Sweep and point and shoot: phonecam-based interactions for large public displays. In: CHI 2005 Extended Abstracts on Human Factors in Computing Systems (CHI EA 2005), pp. 1200–1203. ACM (2005)
8. Zhong, Y., Li, X., Fan, M., Shi, Y.: Doodle space: painting on a public display by cam-phone. In: Proceedings of the 2009 Workshop on Ambient Media Computing (AMC 2009), pp. 13–20. ACM (2009)
9. Jeon, S., Hwang, J., Kim, G.J., Billinghamurst, M.: Interaction with large ubiquitous displays using camera-equipped mobile phones. *Personal Ubiquitous Comput.* 14(2), 83–94 (2010)
10. Boring, S., Baur, D., Butz, A., Gustafson, S., Baudisch, P.: Touch projector: mobile interaction through video. In: Proceedings of the 28th International Conference on Human Factors in Computing Systems (CHI 2010), pp. 2287–2296. ACM (2010)
11. Mistry, P., Maes, P., Chang, L.: WUW - wear Ur world: a wearable gestural interface. In: Proceedings of the 27th International Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA 2009), pp. 4111–4116. ACM (2009)

Head-Free, Remote Gaze Detection System Based on Pupil-Corneal Reflection Method with Using Two Video Cameras – One-Point and Nonlinear Calibrations

Yoshinobu Ebisawa and Kiyotaka Fukumoto

Graduate School of Engineering, Shizuoka University, Hamamatsu, 432-8561 Japan
ebisawa@sys.eng.shizuoka.ac.jp

Abstract. We developed a pupil-corneal reflection method-based gaze detection system, which allows head movements and achieves easy gaze calibration. The proposed gaze detection theory determines gaze points on a PC screen from the vector from the corneal reflection to pupil center, 3D pupil position, two cameras position, etc. In a gaze calibration procedure, after a user is asked to gaze one specific calibration target at a center of a PC screen, the nonlinear characteristic of the eyes has been automatically corrected while the user is using this gaze system. The experimental results show that the proposed calibration method improved the precision of gaze detection during browsing web pages. In addition, the average gaze error in the visual angle is less than 0.6 degree for the nine head positions.

Keywords: Gaze detection, Gaze calibration, Head movement, Pupil.

1 Introduction

For human interface and human behavior monitoring, precise eye-gaze detection with easy calibration procedure is desired. Current commercial gaze detection systems need the gaze more than five points on a PC screen for high precision. However, when the gaze detection system is used for the general public or infants, it is difficult that a user is asked to gaze some points. In previous studies, the calibration methods to gaze a few or no gaze points have been proposed [1][2]. However, their methods include some problems such as a range of user's head movement and easiness for field surveys. Thus, in our previous study, we have developed a gaze detection system based on the pupil-corneal reflection method, which allows large head movements and achieves easy gaze calibration [3].

In this system, an optical system for detecting the pupil and corneal reflection images consist of a camera and a two concentric near-infrared LED ring light source attached to the camera. The inner and outer LED rings generate bright and dark pupil images, respectively. The pupils are detected from the difference image created by subtracting the bright and dark pupil images. Fig. 1 shows the gaze detection theory in

3D space. The 3D coordinates of the pupils are determined by the stereo matching method using two optical systems. The vector from the corneal reflection center to the pupil center in the camera image is replaced by its actual size vector r . The angle between the line of sight and the line passing through the pupil center and the camera (light source) is denoted as θ . The relationship is assumed as $\theta = k|r|$ where k is a constant. The theory allowed head movement of the user and facilitates the gaze calibration procedure. In the automatic calibration method, calibration procedure is accomplished while the user looks around on the PC screen without fixating on any specific calibration target. In the one-point calibration method, the user is asked to fixate on one calibration target at the center of the PC screen in order to correct the innate difference, ΔQ , between the optical and visual axes of the eyeball. In the *two-point calibration* method, in order to correct the nonlinear relationship between θ and $|r|$, which occurs where θ is large, the user is asked to fixate on another target presented at the top of the PC screen as well as the center target. The experimental results show that the three proposed calibration methods improve the precision of gaze detection step by step. In addition, the average gaze error in the visual angle is less than 1 degree for the seven head positions of the user.

However, afterwards, we found a simpler calibration method. In the method, nonlinear relationship between k and r is automatically corrected during the user look around the PC screen area after the one-calibration method is completed. In the present paper, we propose the new calibration and gaze detection methods. The method simultaneously deals with the two problems of the difference, ΔQ , between the optical and visual axes and the nonlinear relationship between θ and $|r|$. In addition, the proposed gaze detection theory became more simple and flexible than our previous theory.

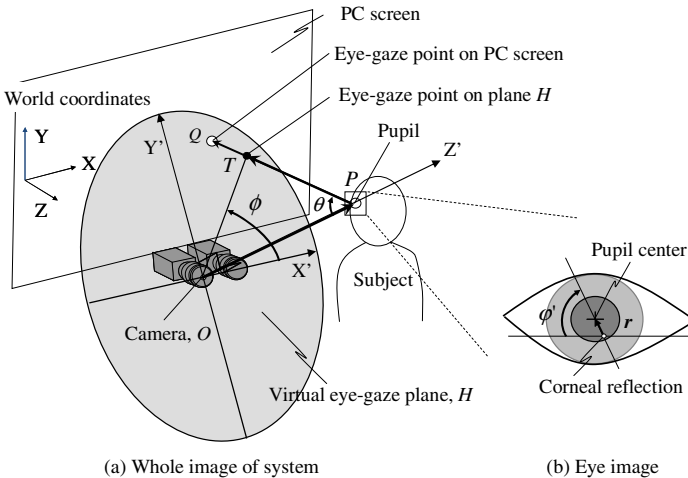


Fig. 1. Gaze detection theory in 3D space

2 Gaze Detection System

2.1 Pupil Detection Principle and Method

In our previous study [3], a light source consisting of near-infrared LEDs that are arranged in two concentric rings (inner and outer rings) was proposed in order to create the bright and dark pupil images, respectively. Since the inner ring is located near the aperture of the camera, the inner ring generates a bright pupil image. Since the outer ring is far from the aperture, the outer ring generates a dark pupil image. In addition, in order to miniaturize the light source, we used the LEDs having two different wavelengths. The structure of the light source is described in the next section.

2.2 System Configuration

Fig. 2 (a) shows an overview of the developed gaze detection system. This system has two optical systems (Fig. 2 (b)), each of which consists of a digital video camera having near-infrared sensitivity, a 16-mm lens, an infrared filter (IR80), and a light source. Light sources consisting of near-infrared LEDs of two different wavelengths that are arranged in two concentric rings (inner: 850 nm, outer: 940nm) are attached to each camera. The pupil becomes brighter in the 850nm ring than the 940nm ring because the transmissivity of the eyeball medium is different. The distance between the LEDs and the aperture of the camera also varies the pupil brightness. The combined effects of the distance and the difference in transmissivity were applied to the light source. An internal synchronization at the hardware level is possible if the cameras are connected to buses of the IEEE-1394 PCI board. The internal synchronization was used to drive two cameras with a slight synchronization difference (670 μ s) because of avoiding mutual light interference of the optical systems.

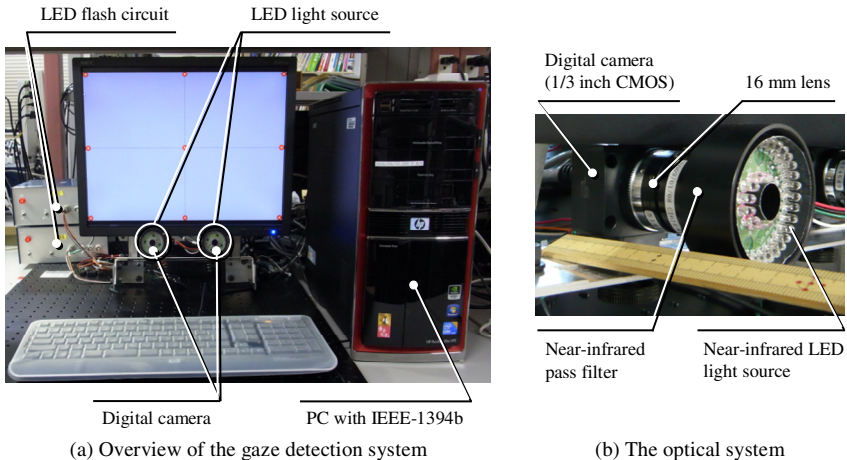


Fig. 2. Overview of the gaze detection system and the optical system

An 8-bit gray scale image of a user's face was input into a personal computer (PC) via the board. The captured image sizes were 640 x 480 pixels. The image was processed using the PC to detect the centers of the pupils and the corneal reflections, which were used to determine the gaze points.

2.3 Image Processing for Detection of the Centers of Pupil and the Corneal Reflection Image

The pupil is detected from the difference image generated from the bright and dark pupil images (Fig. 3 (a)-(c)). The image is processed in the following order: binarization, removal of isolated pixels, noise reduction using mathematical morphology operations, and labeling. The largest and second largest labeled regions were detected as the two pupils. When a pupil was detected in prior frames, the pupil position in the current frame was estimated using a linear Kalman filter, and a small window (70 x 70 pixels) was then applied around the estimated pupil position. On the other hand, when the pupil was not detected in the prior frames (e.g., effect of blinks and eyelashes), the pupils were again searched for in the entire image of the user's face.

The image within the small window is transformed into an image with twice the resolution (140 x 140 pixels) (Fig. 3 (d)). This image from the bright and dark pupil images is processed by binarization and labeling, and an intense and tiny label closest to a center of the double-resolution image is determined as the corneal reflection. The pupil region was again determined by binarizing the difference image, which was obtained after shifting this double-resolution dark pupil image so that the corneal reflection in this dark pupil image may coincide with that in the double-resolution bright pupil image [4]. This process helped to decrease the positional deviation between the bright and dark pupil images while the user's head is moving. Ellipse fitting for the contour of the pupil was then performed. The center of the ellipse was determined as the center of the pupil. The center of gravity considering the values of the pixels in the corneal reflection region in the bright pupil image was determined as the center of the corneal reflection.

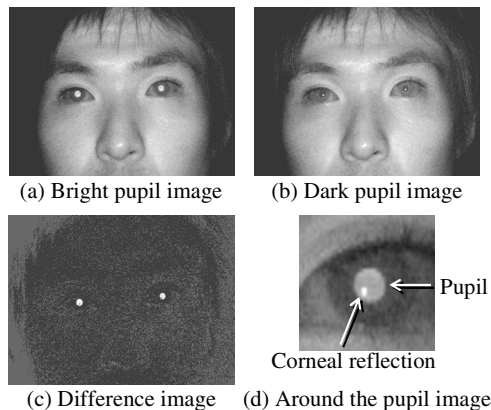


Fig. 3. Detection of pupils and corneal reflection

2.4 Gaze Detection Theory and Calibration Method

In the Fig. 4, O'_1 and O'_2 indicate the pinholes of two stereo-calibrated cameras. We assume that the light source attached to each camera is located at the same position as the corresponding camera. The 3D pupil position P is obtained by the stereo-matching method. The optical axis of an eyeball passes through the pupil P and gaze point Q on the screen plane of the PC display. Now we define the virtual gaze planes (H_1 and H_2) of the cameras for one eyeball. These planes are vertical to the line passing through P and O'_1 or O'_2 , and they pass through O'_1 and O'_2 . The X-axis (X_1 or X_2) of planes H_1 and H_2 is determined as the line intersection between the corresponding plane and the horizontal plane in the world coordinate system ($x - y - z$). H_1 and H_2 rotate according to the displacements of the pupil in the world coordinate system.

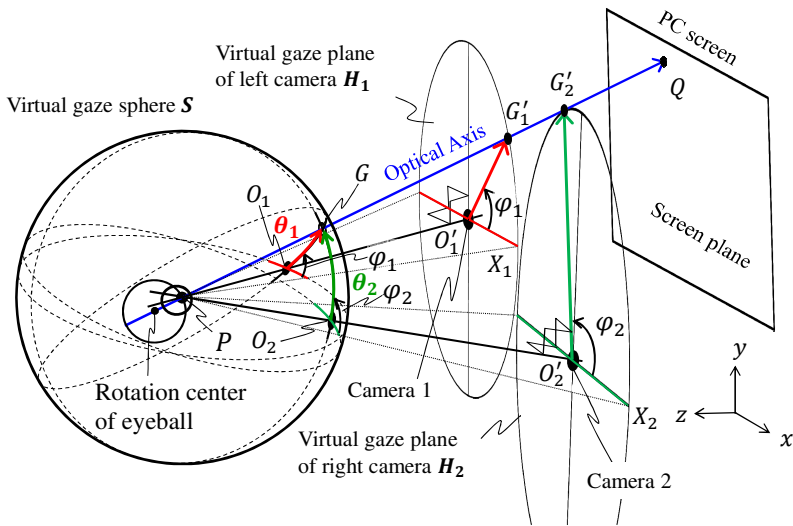


Fig. 4. Gaze detection theory using visual gaze sphere

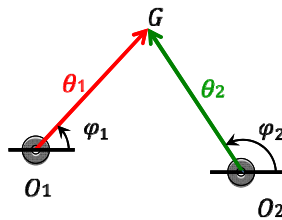


Fig. 5. Angular vectors transformed into 2D coordinate system

Next, we define the virtual gaze sphere S whose center is P . The optical axis PQ has intersection points with sphere S , H_1 and H_2 . The intersection points are denoted as G , G'_1 and G'_2 , respectively. Angular vectors θ_1 and θ_2 on sphere S can be defined

as projections of vectors $O'_1G'_1$ and $O'_2G'_2$ to sphere S . In addition, horizontal axes X_1 and X_2 of planes H_1 and H_2 are projected to sphere S . Here, orientations φ_1 and φ_2 of vectors $O'_1G'_1$ and $O'_2G'_2$ are also projected to sphere S . According to these projections, angular vectors θ_1 and θ_2 on sphere S are transformed into the 2D vectors on a flat plane as shown in **Fig. 4**. Here, vector $\overrightarrow{O_1O_2}$ is expressed as follows:

$$\overrightarrow{O_1O_2} = \theta_1 - \theta_2 \tag{1}$$

Since you can see both relationships $|\theta_1| = \angle O_1PG$ and $|\theta_2| = \angle O_2PG$ referring to Fig. 1, the following equation is obtained.

$$|\overrightarrow{O_1O_2}| = \angle O_1PO_2 \tag{2}$$

Here, we assume that the angular vector θ is a function of r as

$$\theta = f(r). \tag{3}$$

As mentioned before, θ is an angular vector having size θ and orientation φ , and f is monotonically increasing function. In one-point calibration, we assume a linear function as follows:

$$\theta = kr \tag{4}$$

where k is a constant. In general, there is a difference between the optical axis and visual axis. In order to compensate it, measured vector r' is compensated by offset vector r_0 .

$$r = r' - r_0 \tag{5}$$

Accordingly, the following equations are obtained for cameras 1 and 2 from equations (4) and (5).

$$\theta_1 = kr_1 = k(r'_1 - r_0) \tag{6}$$

$$\theta_2 = kr_2 = k(r'_2 - r_0) \tag{7}$$

Here, r'_1 and r'_2 are the pupil-corneal reflection vectors measured from cameras 1 and 2, respectively. r_1 and r_2 are the compensated vectors. From equation (1), (2), (6) and (7), k is given by the following equation.

$$k = \frac{|\theta_1 - \theta_2|}{|r'_1 - r'_2|} = \frac{\angle O_1PO_2}{|r'_1 - r'_2|} \tag{8}$$

Using this value of k , r_0 is determined from equations (6) and (7). Here, r_0 is common for both cameras. Determining the values of k and r_0 means gaze calibration in the *one-point calibration* method. In the *one-point calibration* procedure, the calibrations parameters are determined when a subject fixates on a visual target presented at the center of the PC screen. In the *gaze detection procedure*, first, the pupil-corneal reflection vectors r'_1 and r'_2 are obtained from the images of the two

cameras. By using equations (6) and (7), θ_1 and θ_2 are calculated. Next, the corresponding visual axes are determined from θ_1 , θ_2 and pupil position \mathbf{P} . Finally, the gaze points on the screen are estimated as the intersection points between the screen plane and the visual axis.

In order to compensate the *nonlinear* relationship between θ and $|\mathbf{r}|$, the following equation were used.

$$\theta = f(\mathbf{r}) = g(\mathbf{r})|\mathbf{r}| \quad (9)$$

where $g(\mathbf{r}) = h|\mathbf{r}'|^2 + k_0 \equiv k_f$. Therefore, θ is denoted and calculated by the following equation.

$$\theta = k_f|\mathbf{r}| = (h|\mathbf{r}'|^2 + k_0)|\mathbf{r}' - \mathbf{r}_0| \quad (10)$$

where \mathbf{r}_0 is obtained from the one-point calibration procedure. h and k_0 are constants. These calibration parameters are obtained while a subject is looking around the PC screen *at will*. In order to determine h and k_0 , the relationship between k_f and $|\mathbf{r}'|$ is plotted. The formula of the relationship is obtained by curve fitting. In this *nonlinear calibration*, \mathbf{r}_0 is used for both calibration and gaze detection, as seen in equation (10).

3 Experiments

3.1 Comparison of the Precision of Gaze Detection among the Three Calibration Methods: *one-calibration*, *two-calibration* and *nonlinear calibration*

Ten university students without eyeglasses served as the subjects of this experiment. The distance between the subject's face and the PC screen was approximately 80 cm. In the calibration procedure, the subject fixated on two calibration targets on center and top of the PC screen. The subject was asked to first look at the center calibration target for approximately two seconds, and to then look at the top calibration target. The obtained data for the top target was used only for the *two-point calibration* method [3] but not used for *one-point calibration* method. For *nonlinear calibration*, moreover, the subject was asked to gaze the 25 (five by five) calibration targets arranged on whole the PC screen in order to gaze whole the screen for the subject. The coordinates of the 25 targets were not used for calibration. After the calibration procedure, the subject fixated one by one on the 25 targets for approximately one second in order to compare the precision of gaze detection among the three calibration methods.

Fig. 6 shows the samples of the relationships between k_f and $|\mathbf{r}'|$, which are obtained from the experiment. In our system, the line of sight is obtained from the right and left eyes, respectively. Fig. 7 compares the average precision of the right and left gaze detection among the three calibration methods for each subject. The averages and SDs for *one-calibration* method were a gaze error in visual angle of 0.71 ± 0.41 degrees, whereas those of *two-point calibration* or *nonlinear calibration* were 0.68 ± 0.44 degrees and 0.59 ± 0.30 degrees, respectively.

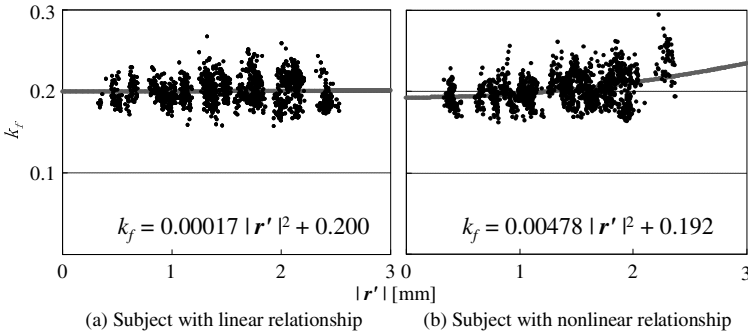


Fig. 6. Relationships between k_f and $|r'|$

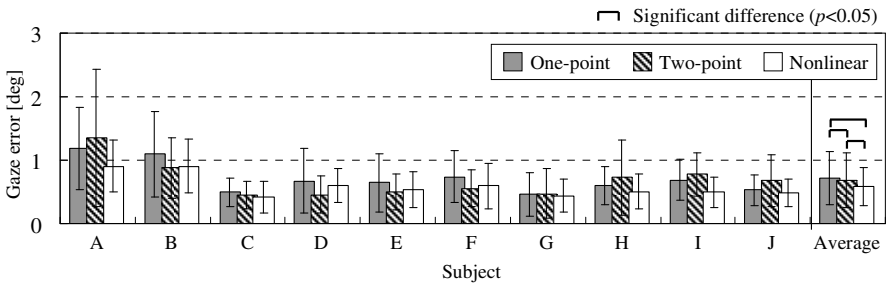


Fig. 7. Comparison of error of the gaze detection of both eyes among the three calibration methods

3.2 Gaze Calibration during Web Browsing

In experiment, for *one-point calibration*, the circle target was used. First the target was large. Suddenly, it shrank. This target attracted the attention of the subject, and made it easy to fixate on the target accurately. After the one-point calibration procedure, the subject was asked to look around the PC screen. During this nonlinear calibration, Google Maps was freely browsed on internet using a PC mouse. After this procedure, the error of gaze points when the subject fixated on the 25 targets was evaluated.

Fig. 8 shows the comparison in the detected gaze point distributions between the one-point and nonlinear calibration methods. You can see that there is a tendency that the nonlinear calibration method shows the errors smaller than the one-point calibration method, especially on the top of the screen. Fig. 9 (a) and (b) show the mean gaze errors distinguishing right and left eyes when each of two subjects fixated on the 25 targets. When the subjects fixated on ten targets presented on the upper part of the screen, it showed clearly difference in gaze error. We conducted another experiment for ten subjects, in which a subject was asked to the 25 targets to investigate the effect of the compensation in the nonlinear calibration method. The results resembled those of the above-mentioned experiment.

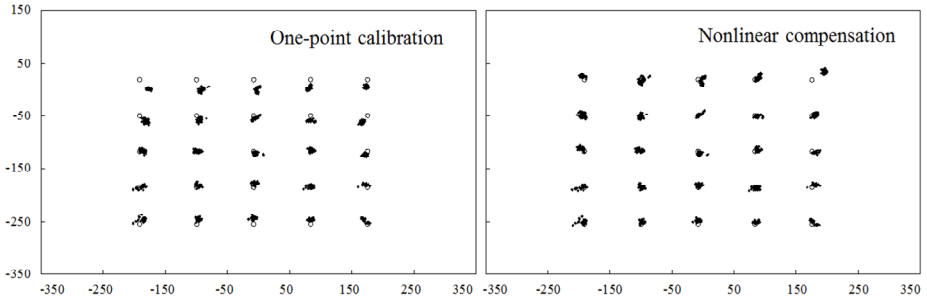


Fig. 8. Detected gaze point distributions between one-point and nonlinear calibration. Circles and dots indicate the target positions and the gaze points, respectively.

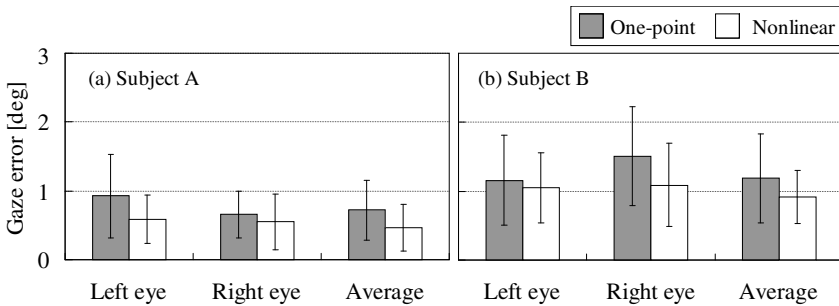


Fig. 9. Comparison in average gaze error between one-point and nonlinear calibration

3.3 Evaluation of Precision by Difference of Head Positions

The subjects were three university students. In the calibration procedure, the subjects were asked to fixate on the same calibration targets as in experiment 2 at approximately 80 cm from the PC screen. After the calibration procedure, the subjects fixated on the same 25 targets as in experiment 2 for the following nine head positions: approximately 70, 75, 80, 85, and 90 cm from the PC screen, and 5 cm to the left, 5 cm to the right, 5 cm to the top and 5 cm to the bottom at 80 cm. The subjects' heads were positioned using a chinrest stand. The calibration values that were obtained at the head position of 80 cm were commonly used for gaze detection at all head positions.

Fig. 10 shows the gaze errors when a subject changed the head position approximately 70, 75, 80, 85, and 90 cm from the PC screen, 5 cm to the left and 5 cm to the right at 80 cm, and 5 cm to the top and 5 cm to the bottom at 80 cm. The nonlinear calibration procedure was conducted at 80 cm. Except for both 5 cm to the top and 5 cm to the bottom, the average gaze error was 0.58 ± 0.33 degrees in the new system while 0.92 ± 0.40 [deg] in the previous system. The new system is improved by 35% in average compared to the previous system. The average gaze error of all head positions was 0.59 ± 0.33 degrees in the new system. The new system uses the digital camera (640 by 480 pixel, 60 frame/sec, non-interlaced scanning) while the previous

used the NTSC analogue camera (640 by 240 field/sec, interlaced scanning). Accordingly, the difference of the cameras may have influenced the precision of gaze detection. However, this result implies that the new calibration method does not reduce the precision of gaze detection.

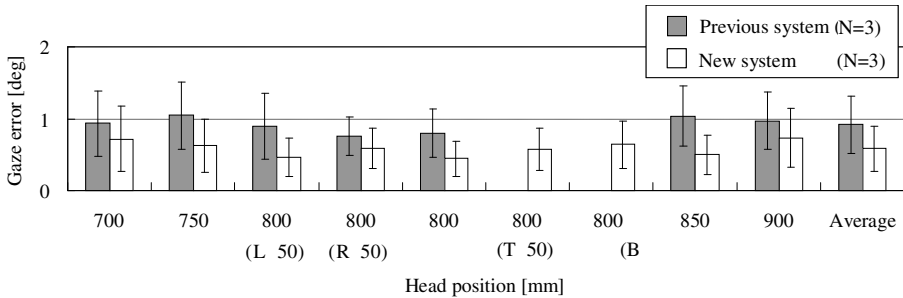


Fig. 10. Comparison in gaze error between our previous and new systems when subject displaced the head position

4 Conclusion

In conclusion, a subject must fixate on one specific target accurately even in the new calibration methods. However, the method improved the nonlinear relationship between θ and $|\mathbf{r}|$ without fixating on any specific position. Looking around in the PC screen is not burden for the subject. Also the new calibration method is very simple. Accordingly, the method is very useful in the experiment, in which the subject is an infant, who is difficult to fixate on a number of points on the screen accurately.

References

1. Model, D., Eizenman, M.: An Automatic Personal Calibration Procedure for Advanced Gaze Estimation Systems. *IEEE Transactions on Biomedical and Engineering* 57(5), 1031–1039 (2010)
2. Villanueva, A., Cabeza, R.: A Novel Gaze Estimation System With One Calibration Point. *IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics* 38(4), 1123–1138 (2008)
3. Ebisawa, Y., Abo, K., Fukumoto, K.: Head-Free, Remote Eye-Gaze Detection System with Easy Calibration Using Stereo-Calibrated Two Video Cameras. In: Stephanidis, C. (ed.) *Posters, Part II, HCI 2011. CCIS*, vol. 174, pp. 151–155. Springer, Heidelberg (2011)
4. Ebisawa, Y., Nakashima, A.: Increasing Precision of Pupil Position Detection Using the Corneal Reflection. *The Institute of Image Information and Television Engineers* 62(7), 1122–1126 (2008)

Design and Usability Analysis of Gesture-Based Control for Common Desktop Tasks

Farzin Farhadi-Niaki¹, S. Ali Etemad², and Ali Arya³

¹ School of Electrical Engineering and Computer Science, University of Ottawa,
Ottawa, Canada

ffarh101@uottawa.ca

² Department of Systems and Computer Engineering, Carleton University, Ottawa, Canada

ali_etemad@carleton.ca

³ School of Information Technology, Carleton University, Ottawa, Canada

arya@carleton.ca

Abstract. We have designed and implemented a vision-based system capable of interacting with user's natural arm and finger gestures. Using depth-based vision has reduced the effect of ambient disturbances such as noise and lighting condition. Various arm and finger gestures are designed and a system capable of detection and classification of gestures is developed and implemented. Finally the gesture recognition routine is linked to a simplified desktop for usability and human factor studies. Several factors such as precision, efficiency, ease-of-use, pleasure, fatigue, naturalness, and overall satisfaction are investigated in detail. Through different simple and complex tasks, it is concluded that finger-based inputs are superior to arm-based ones in the long run. Furthermore, it is shown that arm gestures cause more fatigue and appear less natural than finger gestures. However, factors such as time, overall satisfaction, and easiness were not affected by selecting one over the other.

Keywords: Usability study, human factors, arm/finger gestures, WIMP.

1 Introduction

The new wave of input systems in video game consoles (such as Nintendo Wii, Xbox Kinect, and PlayStation Move) is leading the new generation of Human-Computer Interaction (HCI) systems to focus on creating interfaces that are more intuitive and user-friendly. While the gaming industry is currently leading the way using the aforementioned consoles, it will not be long before users will be controlling advanced and simple Virtual Reality (VR) and computer systems using body gestures that feel intuitive. Having an HCI system designed intuitively, thus, can provide higher user satisfaction and better performance. Development of reliable gesture recognition algorithms, choosing appropriate gestures, and studying the usability of these gesture-based methods are among topics that require the attention of researchers.

In this paper, we describe the design and implementation of a system capable of interacting with natural gesture inputs through computer vision methods. The vision

sub-system is based on the Kinect depth-based camera. Recognition algorithms have been studied and implemented to form a robust system. The operating environment is a simulated computer desktop containing several objects such as windows and icons that simplifies the user interface and allows better control over test sessions. Gestures, both using full arm and only using fingers, have been carefully designed and assigned to replace mouse inputs for common desktop tasks. Successive to implementation, the system is tested with multiple users which provide the feedback needed to analyze the usability of such systems with respect to factors such as precision, efficiency, ease-of-use, fun-to-use, fatigue, naturalness, and overall satisfaction.

The major contributions of this study are: a) choice of natural gestures, b) usability study for gesture-based input, and c) system design (UI and gesture recognition) and relatively novel use of existing API's to implement gesture recognition method. This method conserves the developing time (no need for making samples and perform training and testing sessions) and running time for gesture recognition and user interaction compared to learning-based traditional method.

2 Related Work

Employing natural arm and finger gestures for VR applications can be studied in two different aspects: technical design and implementation, and usability. Through the following we review some relevant vision-based gesture systems in the two mentioned fields.

Detecting hand gestures has been subject to extensive research. Hidden Markov models (HMM) are one of the popular classifiers for this purpose. Marcel et al. [1] employed input-output HMMs for tracking variations in the skin color of the human body. Similarly, Chen et al. [2] employed HMMs for detecting hand postures. The AdaBoost algorithm was revised and used by Liu et al. [3] to automatically recognize users' hand from the video stream. Yu et al. [4], proposed a hand gesture feature extraction method using multi-layer perceptrons. Raheja et al. [5] used principal component analysis (PCA) for hand pattern matching. Other techniques such as cascade classifiers often used for face tracking applications have also been utilized to recognize hands and various parts of the human body [6].

The second parameter in need for in depth study, as mentioned earlier, is the usability aspect of natural arm and finger gestures for practical and VR system inputs. In this regard, Cabral et al. [7] discussed numerous issues associated with the use of gestures as input modes. Their studies showed that both time and fatigue increases when gestures are used for simple pointing tasks. Villaroman et al. [8] show that Kinect-assisted instruction can be utilized to accomplish certain learning results in HCI courses. Moreover, through their study, it is confirmed that OpenNI, a system that is also employed in this research, is a reliable and effective tool to be utilized along with Kinect. It was shown that when the two are used together, students are provided with a hands-on experience on gesture based natural user interaction systems and technologies. Through another study on using Kinect for VR interaction, Kang et al. [9] used distance information and joints' location information and achieved higher recognition

rates. They also showed that their system was 27% faster than the mouse device. Bragdon et al. [10] developed a system that combines touch and air gesture hybrid interactions for small developer group meetings. Their proposed system proves applicable with different devices such as multi-touch screens, mobile touch devices, and Kinect. The use and usability of hand gestures for tasks such as making telephone calls, operating the television, and executing mathematical calculations has been studied by Bhuiyan and Picking [11]. Their study suggests that such technologies can benefit the elderly and the disabled users by causing more independence while some challenges still remain to overcome. Applications of gestures as inputs for medical systems have also been explored. In a study conducted by Ebert et al. [12] rebuilding images from a CT data was tested and it was shown that image recreation time using gestures was longer than using mouse/keyboard. The system, however, maintained certain advantages such as reducing the potential for infection, for both patients and staff.

The provided review on the relevant literature shows that the use of natural arm/hand/finger gestures for interaction with different systems has been growing. This trend is especially increasing as new generation game consoles such as Kinect are becoming more available and convenient to purchase and develop by researchers. Free developing and computer vision tools such as OpenNI and OpenCV also aid and accelerate the process.

3 Methodology

In this section we describe the design and methodology used for developing our gesture-based simulated desktop interaction system. The three different steps of the system design are described through the following sections: user interface design, natural gesture selection, and gesture recognition module.

3.1 User Interface Design

The simulated desktop for the system was developed using the Allegro library. Allegro is an open source library used for game and multimedia programming. Its cross-platform nature makes it easy to integrate with other modules of the system. The Post-WIMP (windows-icons-menus-pointers) [13] design is adapted for the desktop while neutral colors are utilized to reduce user error or bias. Novice users can learn WIMP user interfaces easily, as they are very good at abstracting workplaces due to their analogous paradigm to documents like paper sheets or folders. Having a rectangular region on a 2D flat screen makes them preferable to system developers while their generality also makes them a good fit in multitasking environments.

3.2 Natural Gesture Selection

We first studied several possible natural gestures suitable for a Post-WIMP user interface [14] [15], and then defined the best matches of the predefined gestures to our

prototype. One determining criterion for the selected gestures is the intuitive naturalness of the actions and effects. Tables 1 and 2 present the final selected arm and finger gestures for the proposed system and corresponding descriptions. Table 3 presents the analogies for the three input mechanisms with respect to one another and corresponding actions.

Table 1. Final design for arm gesture set



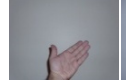
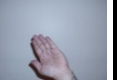

Description	Hand pushing		Hand moving		Hand circling
Hand gestures					

Table 2. Final design for finger gesture set


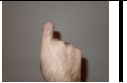


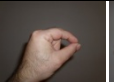

Description	Finger tapping		Finger moving		Pinching	
Finger gestures						

Table 3. Arm and finger gestures, and mouse analogies

Finger gestures	Arm gestures	Mouse	Actions
Tapping	Pushing	Left click	Selecting/Opening/Closing/Dropping
Moving	Moving	Moving	Pointer movement
Pinching	Circling	Drag	Grabbing/Resizing

3.3 Gesture Recognition Module

For arm gesture detection and recognition, some predefined features (circle and push) of applied APIs (OpenNI and NITE) are utilized, and a very efficient and accurate system is developed. Such pre-defined functionalities, however, do not exist for finger gestures (pinch and tap) detection and classification in OpenNI and NITE. Therefore we have implemented the algorithm in OpenCV (Fig. 1) to detect the fingers (Fig. 2).

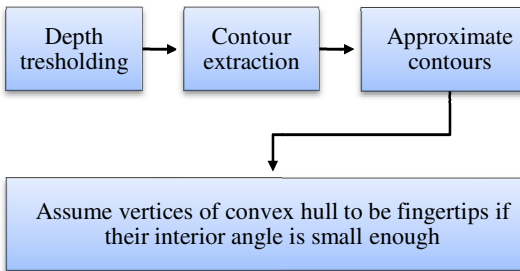


Fig. 1. Fingertip detection algorithm

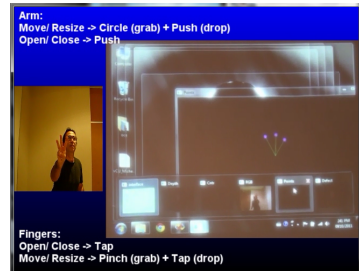


Fig. 2. Finger detection outputs

The “tapping” gesture has been defined based on the depth change of fingertip (z_p) comparing to the depth of hand/palm’s centre (z_h), and a proper threshold (D) as shown below:

$$\text{Tapping happens when: } |\mathbf{z}_h - \mathbf{z}_p| > D \quad (1)$$

The “pinching” gesture has been defined based on the distance between the point finger’s tip (x_p, y_p, z_p) and thumb’s tip (x_t, y_t, z_t) based on the following:

$$\text{Pinching happens when: } \begin{cases} |x_p - x_t| < \varepsilon \\ |y_p - y_t| < \varepsilon \\ |z_p - z_t| < \varepsilon \end{cases} \quad (2)$$

We have designed an Algorithm (similar to our previous work for arm gestures in [16]) to control our user interface objects utilizing the recognized finger/arm gestures (circle and push are replaced by pinch and tap). This algorithm also recognizes index finger and thumb, with a possibility of orderly detecting all five fingers.

4 User Experiments

In this study two different interactive variables namely arm and finger gestures are employed and compared. A set of usability parameters are analyzed for each input method when performing combined tasks with two difficulty levels (simple and complex), on big-screen display. According to our previous study (gestures vs. mouse) [16], using gestures on big-screen was proved to be superior to using gestures on desktop-screen. Therefore, we have chosen solely big-screen display for this inter-gestures study. The usability parameters consist of human factors such as ease of use, fatigue, naturalness, pleasantness, and overall satisfaction, as well as performance factors such as efficiency and effectiveness.

4.1 Training Session

In this session participants are asked to practice primitive tasks for a period of 30 minutes in order to get acquainted with the system prior to participating in the main test. Furthermore, this phase plays a major role in validity of particular satisfaction criteria such as fatigue and naturalness.

4.2 Test Session

During the test session, the different variety of tasks using each input method on big-screen is examined. As described earlier, the two variables, input method and task difficulty, combine to generate four states. Each state is examined independently for each participant. Figure 3 presents snapshots of some activities performed during the test sessions.

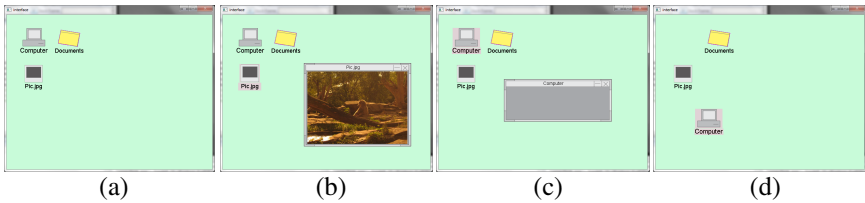


Fig. 3. User interface: (a) initial configuration, (b) object opened, (c) object resized, and (d) object moved

4.3 Questionnaire and Observations

Following the test sessions, participants are asked to provide their feedback through a questionnaire from which user satisfaction criteria are extracted. Ratings are scaled from 1 to 5 (1 for absolutely unsatisfied and 5 for extremely satisfied). Extra written feedback is also acquired for further information regarding both participants and system. During the different states of the test sessions, time and error are observed and recorded for each user.

5 Results and Discussions

This study is conducted using 10 participants (5 males and 5 females) and in the age range of 26 to 36 (average of 30 years old).

5.1 Hypotheses and Analyses

For the different factors being studied, two-way repeated analysis of variances (ANOVA) is carried out for two independent variables:

1. Difficulty (simple task vs. complex task)
2. Input method (finger gestures vs. arm gestures)

All experiments were carried out on the big-screen and at $p < 0.05$ significance level and for 10 participants.

Notation: In our analyses, we calculate the mean and standard deviation for different variables in the forms of M_{variable} (e.g. M_{simple} is the mean for simple task) and SD_{variable} (e.g. SD_{finger} is the standard deviation for finger gestures). Moreover, $F(\text{df}, \text{MS})$ is the test statistic (F-ratio) in which df and MS are the degree of freedom and mean square respectively for the variables (within variables when more than one, and within subjects). The F-ratio is calculated using $MS_{\text{variable(s)}}/MS_{\text{error(s)}}$ and P is the probability value.

Table 4 presents our statistical analyses, hypotheses, and results for different factors.

Table 4. ANOVA analyses, hypotheses, and results for different factors

	Hypotheses	Variable 1	Variable 2	Variables 1 & 2	Results
Time	Using finger gestures is faster than using arm gestures as inputs.	$F(1,617.01) = 202.5$ P = 0.00	$F(1,2.86) = 0.31$ P = 0.59	$F(1,0.13) = 0.05$ P = 0.82	<u>Rejected</u>
		($M_{\text{simple}} = 13.35$ $SD_{\text{simple}} = 2.30$) vs. ($M_{\text{complex}} = 21.21$ $SD_{\text{complex}} = 3.43$)			
Easiness	Using finger gestures is easier than using arm gestures as inputs.	$F(1,0) = 0$ P = 1	$F(1,0) = 0$ P = 1	$F(1,0.10) = 2.25$ P = 0.16	<u>Rejected</u>
Fatigue	Using finger gestures causes less fatigue than using arm gestures as inputs.	$F(1,0) = 0$ P = 1	$F(1,4.90) = 12.25$ P = 0.00	$F(1,0) = 0$ P = 1	<u>Confirmed</u>
			($M_{\text{finger}} = 4.50$ $SD_{\text{finger}} = 0.60$) vs. ($M_{\text{arm}} = 3.80$ $SD_{\text{arm}} = 0.69$)		
Naturalness	Using finger gestures is more natural than using arm gestures as inputs.	$F(1,0.02) = 1.00$ P = 0.34	$F(1,11.02) = 441.0$ P = 0.00	$F(1,0.02) = 1.00$ P = 0.34	<u>Confirmed</u>
			($M_{\text{finger}} = 5$ $SD_{\text{finger}} = 0$) vs. ($M_{\text{arm}} = 3.95$ $SD_{\text{arm}} = 0.22$)		
Pleasantness	Using finger gestures is more pleasant than using arm gestures as inputs.	$F(1,0.40) = 6.00$ P = 0.03	$F(1,0) = 0$ P = 1	$F(1,0.40) = 6.00$ P = 0.03	<u>Rejected</u>
		($M_{\text{simple}} = 4.80$ $SD_{\text{simple}} = 0.41$) vs. ($M_{\text{complex}} = 4.60$ $SD_{\text{complex}} = 0.50$)		($M_{\text{finger-simple}} = 4.90$ $SD_{\text{finger-simple}} = 0.31$) vs. ($M_{\text{finger-complex}} = 4.50$ $SD_{\text{finger-complex}} = 0.52$)	
Overall Satisfaction	Overall, using finger gestures as inputs is a more popular experience compared to arm gestures.	$F(1,0.40) = 3.27$ P = 0.10	$F(1,0.40) = 3.27$ P = 0.10	$F(1,0.10) = 2.25$ P = 0.16	<u>Rejected</u>

Figure 4 shows the times taken to complete simple and complex tasks using arm and finger gestures. Analysis of users’ feedback regarding the primitive tasks is shown in Fig. 5, where pinching to resize and pushing to close a window were the most difficult gestures, due to the relatively small control access area of the objects.

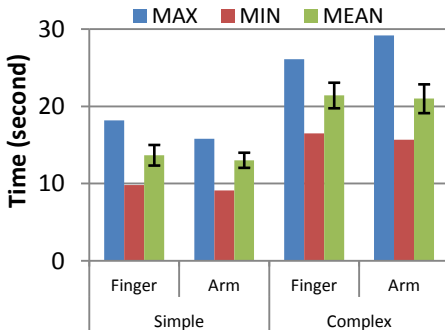


Fig. 4. Temporal statistical factors

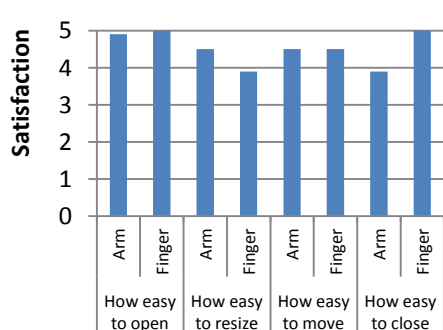


Fig. 5. Satisfaction on primitive tasks

As the following figures show, the Tapping was the smoothest gesture with the least errors (average number of trials) in both simple and complex tasks, while the Circling had the highest error in the simple task. However, the most errors happened with the Pushing during the complex task on the action of closing a window.

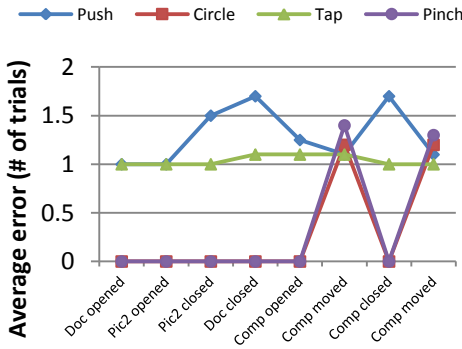


Fig. 6. Gestures errors in complex task

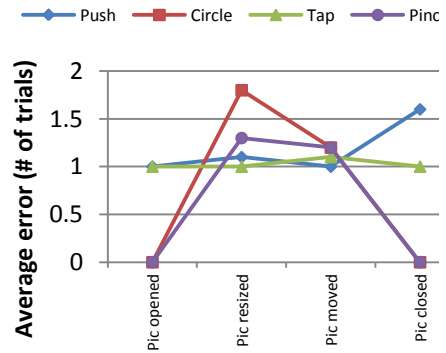


Fig. 7. Gestures errors in simple task

5.2 Discussion

Before disclosing the defined gestures to the participants, we asked them to try grabbing an object (here an icon) naturally and based on their common sense. Interestingly, 85% of participants in their first guess could correctly pick an object on screen by Pinching gesture. This anecdotal evidence of natural grabbing indicates that we have been successful in defining our finger gestures in a natural way.

This study is a supplementary work to the authors’ previous research, comparing the arm gestures to mouse/keyboard [16], using the same user interface. The results in [16] are summarized as follows:

The gesture inputs are significantly slower and more fatiguing than using a mouse. Moreover, using a mouse is significantly easier than using arm gestures while neither inputs hold a significant popularity over the other. For the naturalness and the pleasure factors, the arm gestures as inputs do not feel significantly more natural or more fun to use compared to mouse. However, it is revealed that using arm gestures on big-screen is significantly more natural and more pleasant than using a mouse on both the desktop and the big-screen. Also it is shown that arm gestures used on big-screen is significantly more pleasant compared to when it is used on desktop.

According to the provided statistical analyses in the present study, we summarize our hypotheses verification as follows:

In general, the main result of this experiment is that fatigue is less for finger compared to arm gestures. To elaborate on, the naturalness and the fatigue factors analyses support our initial hypotheses, meaning the finger gestures significantly are more natural and cause less fatigue as inputs compared to the arm gestures. The initial hypotheses in terms of time, overall satisfactory, and easiness are rejected, implying that finger and arm maintain similar performances and popularities among participants, and

neither finger gestures nor arm gestures are significantly easier than the other. Moreover, for the pleasure factor, the initial hypothesis is rejected as well, meaning the finger gestures compared to the arm gestures are not significantly more pleasant to employ as inputs. However, it is revealed that finger gestures are significantly more pleasant for simple tasks rather than complex ones.

Finally, through written feedback, most participants preferred “mostly finger” as their preferred combination of using arm and/or finger gestures, which is in positive correlation with the findings of this study.

As shown in Fig. 8, using arm is easier in short term (simple tasks). However, it is easier to use finger in the long run (complex tasks). In addition, using finger in the short term is the most pleasant, and in the long term is the least pleasant. The overall satisfaction had its highest level on the simple task using finger gestures, and its lowest level on the complex task using arm gestures.

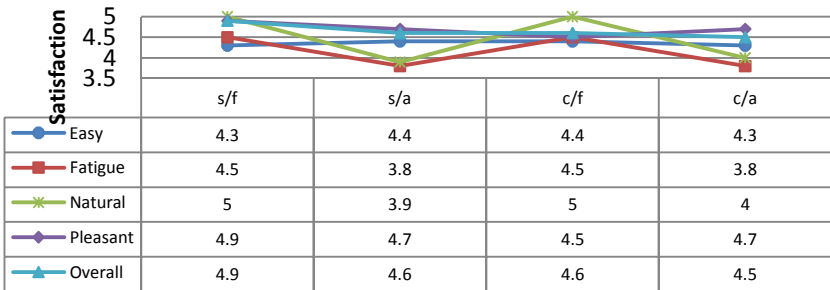


Fig. 8. Satisfaction comparison (s: simple, c: complex, f: finger, a: arm)

6 Conclusion

Using Kinect depth-based cameras along with OpenNI, NITE, and OpenCV, a gesture detection and recognition system has been developed and linked to a simulated desktop environment. Gesture studies were carried out and natural and intuitive gestures were chosen and utilized for performing various tasks in the environment. Two sets of gestures were designed, one using the arm and one using fingers. The performed tasks were designed with different difficulty levels for extraction of more information regarding the system at hand.

A comprehensive usability study indicated that finger-based gestures appeared more natural and less tiring for participants. However, this variable showed no significant effect on time, easiness, and overall satisfaction. Finally, through written feedback, most participants indicated that they would prefer a combination of fingers and arm gestures with “mostly fingers” as their method of choice for such applications.

The findings of this study, we believe, can be widely used for designing gesture-based systems, especially for WIMP interfaces. In general, finger gestures would be preferred, especially for longer lasting application which can cause more fatigue. For more rare functionalities, however, arm gestures would also be a valid choice.

References

1. Marcel, S., Bernier, O., Viallet, J.E., Collobert, D.: Hand Gesture Recognition Using Input-Output Hidden Markov Models. In: Conference on Automatic Face and Gesture Recognition (2000)
2. Chen, F., Fu, C., Huang, C.: Hand Gesture Recognition Using a Real-Time Tracking Method and Hidden Markov Models. *Image and Vision Computing*, 745–758 (2003)
3. Liu, Y., Zhang, P.: Vision-Based Human-Computer System Using Hand Gestures. *International Conference on Computational Intelligence and Security*, 529-532 (2009)
4. Yu, C., Wang, X., Huang, H., Shen, J., Wu, K.: Vision-Based Hand Gesture Recognition Using Combinational Features. In: Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, pp. 543–546 (2010)
5. Raheja, J.L., Shyam, R., Kumar, U., Prasad, P.B.: Real-Time Robotic Hand Control Using Hand Gestures. In: Second International Conference on Machine Learning and Computing, Bangalore, India, pp. 12–16 (2010)
6. Chen, Q., Cordea, M.D., Petriu, E.M., Varkonyi-Koczy, A.R., Whalen, T.E.: Human-Computer Interaction for Smart Environment Applications Using Hand-Gesture and Facial-Expressions. *International Journal of Advanced Media and Communication* 3(1/2), 95–109 (2009)
7. Cabral, M.C., Morimoto, C.H., Zuffo, M.K.: On the Usability of Gesture Interfaces in Virtual Reality Environments. In: Proc. of the Latin American Conf. on Human-Computer Interaction, Cuernavaca, Mexico, pp. 100–108 (2005)
8. Villaroman, N., Rowe, D., Swan, B.: Teaching Natural User Interaction Using OpenNI and the Microsoft Kinect Sensor. In: Proc. of the 2011 Conference on Information Technology Education, West Point, New York, USA, pp. 227–232 (2011)
9. Kang, J., Seo, D., Jung, D.: A Study on the Control Method of 3-Dimensional Space Application Using KINECT System. *International Journal of Computer Science and Network Security* (2011)
10. Bragdon, A., DeLine, R., Hinckley, K., Morris, M.R.: Code Space: Touch+Air Gesture Hybrid Interactions for Supporting Developer Meetings. In: Proc. of ACM International Conference on Interactive Tabletops and Surfaces, Kobe, Japan (2011)
11. Bhuiyan, M., Picking, R.: A Gesture Controlled User Interface for Inclusive Design and Evaluative Study of Its Usability. *Journal of Software Engineering and Applications* (2011)
12. Ebert, L.C., Hatch, G., Ampanozi, G., Thali, M.J., Ross, S.: *You Can't Touch This: Touch-free Navigation Through Radiological Images*. SAGE Publications (2011)
13. Van Dam, A.: Post-WIMP User Interfaces. *Communications of the ACM* 40(2), 63–67 (1997)
14. Vatavu, R., Pentiu, S., Chaillou, C.: On Natural Gestures for Interacting in Virtual Environments. *Advances in Electrical and Computer Engineering* 5(24), 72–79 (2005)
15. Pavlovic, V., Huang, T.: Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19(7), 667–693 (1997)
16. Farhadi-Niaki, F., GhasemAghaei, R., Arya, A.: Empirical Study of a Vision-based Depth-Sensitive Human-Computer Interaction System. In: 10th Asia Pacific Conference on Computer Human Interaction, Matsue, Japan, pp. 101–108 (2012)

Study of Eye-Glance Input Interface

Dekun Gao, Naoaki Itakura, Tota Mizuno, and Kazuyuki Mito

The University of Electro-Communications,
1-5-1 Chofugaoka, Chofu, Tokyo 182-8585, Japan
{gaodekun,mizuno}@uec.ac.jp, {ita,mito}@se.uec.ac.jp

Abstract. Optical measurement devices for eye movements are generally expensive and it is often necessary to restrict user head movements when various eye-gaze input interfaces are used. Previously, we proposed a novel eye-gesture input interface that utilized electrooculography amplified via an AC coupling that does not require a head mounted display[1]. Instead, combinations of eye-gaze displacement direction were used as the selection criteria. When used, this interface showed a success rate approximately 97.2%, but it was necessary for the user to declare his or her intention to perform an eye gesture by blinking or pressing an enter key. In this paper, we propose a novel eye-glance input interface that can consistently recognize glance behavior without a prior declaration, and provide a decision algorithm that we believe is suitable for eye-glance input interfaces such as small smartphone screens. In experiments using our improved eye-glance input interface, we achieved a detection rate of approximately 93% and a direction determination success rate of approximately 79.3%. A smartphone screen design for use with the eye-glance input interface is also proposed.

Keywords: Eye gesture, eye-glance, AC-EOG, smartphone, Screen design.

1 Introduction

Human computer interactions (HCI) are an important field in computer science. Many computer interfaces are being developed for people with disabilities[2-3]. One such computer interface type utilizes eye-gaze behavior. An eye-gaze interface can be faster than a computer mouse for inputting user selections and is convenient in situations where it is essential that a user keeps his or her hands free in order to perform other tasks [4-5]. Previously, eye-gaze interfaces with direct input methods that rely on detecting the user's gaze point have been most frequently studied[1-5]. However, optical measurement devices for eye movements are generally expensive, and it is often necessary to restrict the user's head movements when using various eye-gaze input interfaces to prevent the introduction of diagonal eye movements.

Based on the amplification method used, there are currently two types of electrooculographs (EOGs) in use: DC coupled (DC-EOG) and AC coupled (AC-EOG). In DC-EOG, voltage must be applied manually to the amplifier in order to adjust the baseline to zero in response to changes in the resting potential. In contrast, AC-EOG does not require such adjustments.

In our earlier studies [6-9], we discussed an eye-gaze input interface that combined a head mounted display (HMD) with an AC-EOG. That interface was relatively inexpensive and enabled users to escape head movement restrictions. It was also noteworthy because it permitted the introduction and use of diagonal eye movements, thus introducing 12 possible eye-gaze movement choices. However, it was considered suboptimal because it required the use of a costly HMD that was time consuming to put on and troublesome for the users. In addition, that particular gaze input method required the user to spend uncomfortable amounts of time watching the target.

Separately, another eye-gaze interface that can be used to control a cursor by detecting the user's gaze direction has been investigated[10-12]. However, the number of movement directions that can be identified by this interface is between 4 and 8 without target. Furthermore, taking into consideration the original function of an eye, it is reasonable to assume that allowing users to look directly at the target when detecting eye-gaze movements would allow for more natural eye movements.

In the study described in Reference 8, a display based on measuring diagonal eye movements was proposed. In that interface, to determine the eye's position during diagonal motion, information about vertical eye movements is combined with the horizontal AC-EOG signals. This design permitted 12 possible choices, the mean accuracies of which were 89.2%.

This interface determines the choices made from the vertical eye movement direction and amount of eye movement in the horizontal direction. Thus, to obtain a precise measurement of the amount of eye movement in the horizontal direction, the relative position of the input screen and the eyeball must be known, which made it necessary to secure the HMD to the user's head.

2 Eye Gesture Input Interface

In Reference 1, we reviewed previous input methods with an aim towards reducing system costs and restrictions placed on the user. The result was a novel eye-gesture input interface that did not require a HMD. Instead, direction combinations for eye-gaze displacement were used as the selection method. We found that eye-gaze displacements could be determined precisely using a derivative EOG signal amplified via AC coupling. A desktop display design created for use with the eye-gesture input interface is shown in Fig. 1. In that study, it was assumed that eye-gesture movements followed oblique patterns (upper left, lower left, upper right, lower right), and that each pattern consisted of a combination of two movements.

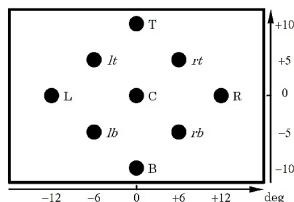


Fig. 1. Experimental screen design for eye-gesture input interface

Users were instructed to initiate use of the input method by following the three instructions provided below:

1. Glance at the central C while pressing the enter key (Gesture start).
2. Glance first at the targets lt, lb, rt, and rb.
3. Glance at the targets C, T, B, L, and R (Gesture finish).

Using this method, the eye movements of the pattern could be saved as a combination of $4 \times 3 = 12$ pairs. For this interface, eye movement was detected twice in the oblique direction during steps 1~3. Furthermore, the user eye-gesture input is determined as a pattern combination. The advantage of this eye-gesture input interface is that determinations were made from the combination of diagonal eye movements, and there was no need to calculate precisely the amount of eye movement and the like.

Accordingly, the need to fix the relative position of the eyeball and the input screen is eliminated, which also eliminates the need for a HMD, and an input interface that can be utilized in numerous situations was made possible.

However, this interface required the user to declare the initiation of an eye gesture by, for example, blinking or pressing enter key. In this paper, we propose an upgraded eye-glance input interface that can be used consistently, and which does not require the user to declare his or her initiation of an eye gesture. A decision algorithm created for use with the eye-glance input interface is also proposed.

3 Eye-Glance Input Interface

3.1 Eye-glance

An eye-glance is defined as an action used to obtain a characteristic waveform whereby a user glances at a target for just a moment. It is different from reading and/or searching text on a screen.

For a small screen, such as a smartphone, we restricted recognition to four possible choices – from the center of the screen to the four corners in a round-trip oblique series of eye movements (upper left, lower left, upper right, lower right) as shown in Fig. 2.

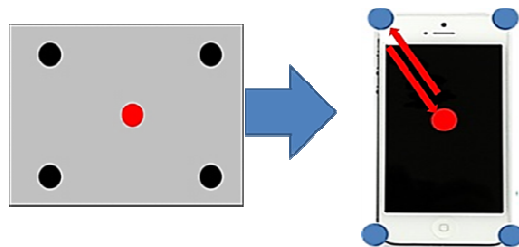


Fig. 2. Eye-glance input interface using a smartphone screen (target and arrow aren't shown)

An eye-glance has the following three characteristics:

1. A dual eye movement from and to the center is provided to the horizontal AC-EOG, as shown in Fig. 3.

2. There is a “pause” time τ between the times of occurrence of the dual eye movement.
3. The vertical AC-EOG is the same as the horizontal AC-EOG, and occurs at the same time.

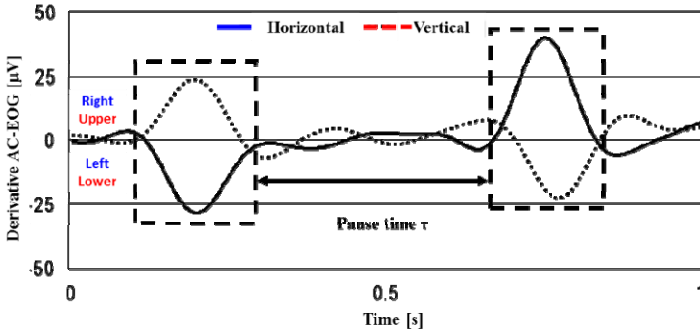


Fig. 3. Example of a derivative AC-EOG signal created from a single eye-glance

With these characteristics in mind, we performed experiments to evaluate a standard that can be used to distinguish eye-glance input from normal eye movement.

3.2 Methods

Four healthy male subjects ranging in age from 21 to 24 years old (mean = 22.5 years, SD = 1.3 years) volunteered to participate in our study. None of the subjects reported any medical or psychiatric problems at the time of testing. All volunteers were informed about the aims and the possible risks of the study. During the experiments, the subjects were asked to sit on a chair and manipulate a keyboard with their right hand. The AC-EOG signals were captured by 10-mm Ag-AgCl metal electrodes (NIHON KODEN Corp., Tokyo, Japan) placed around the subjects’ eyes, as shown in Fig. 4. For the horizontal AC-EOG, two electrodes were placed 2.0 cm lateral to the outer canthi. For the vertical AC-EOG averaging the two vertical AC-EOGs, four electrodes were placed 2.0 cm above and 2.0 cm below the two electrodes for the horizontal AC-EOG. Finally, another electrode placed on the left ear served as a ground.

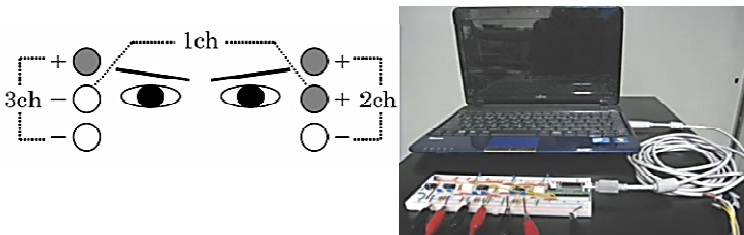


Fig. 4. Electrodes, channels, and our original apparatus

The AC-EOG signals obtained from the amplifiers were fed into an apparatus of our own design connected to a PC[1]. The amplifiers have a gain of 10,000, a high pass analog filter with a 0.1 Hz cut-off frequency, and a low pass analog filter with a 10.0 Hz cut-off frequency. The amplified AC-EOG signals were sampled at a rate of 100 Hz using a 12-bit resolution. Before conducting the experiments, the following assumptions and preconditions were set:

- The average value of 2-ch and 3-ch was used to create a fourth channel (4-ch) on the computer
- 1-ch corresponds to the horizontal AC-EOG
- 4-ch corresponds to the vertical AC-EOG
- Positive and negative values for 1-ch correspond to eye movements to the left and right
- Positive and negative values for 4-ch correspond to upward and downward eye movements

3.3 Experimental Procedure

Each test subject performed a total of 10 eye-glance attempts during which he glanced at each of the four corners. A standard value for a 6° movement was used to calibrate the interface prior to the experiments. Each attempt was performed using the following steps:

1. The subject picks up and uses a smartphone normally (Free time).
2. The eye-glance maneuver begins when the subject focuses his gaze on the center of the screen and then presses a timer.
3. The subject then performs a round-trip series of eye movements in an oblique direction from and to the center screen, viewing each corner in series.
4. The eye-glance maneuver finishes when subject return his gaze to the center of the screen a final time and presses the timer again.
5. The subject then operates the smartphone normally (Free time).

3.4 Displacement Calculation

The derivative AC-EOG signal, an example of which is shown in Fig. 5, was used to calculate the displacements of the eye movements. The value of the derivative AC-EOG is directly proportional to the velocity of the eye movement. A saccade is defined as a rapid eye movement, during which the eye's velocity changes from zero to a large value and then returns to zero. Therefore, a 0-to-0 interval of the derivative AC-EOG (0-0 wave) such as that shown as h1 in Fig. 5, can be considered a saccade interval.

In previous studies[9], the integral of a 0-0 wave is used as the proportional value to the displacement of a saccade. However, the maximum amplitude of a 0-0 wave is also thought to have a value that is proportional to the displacement of a saccade.

In this study, the large maximum amplitude of a saccade can be detected from the maximum amplitude value of a 0-0 wave. In contrast, a state of prolonged fixation is considered an eye-gaze state. Whenever an eye-gaze state was detected, the

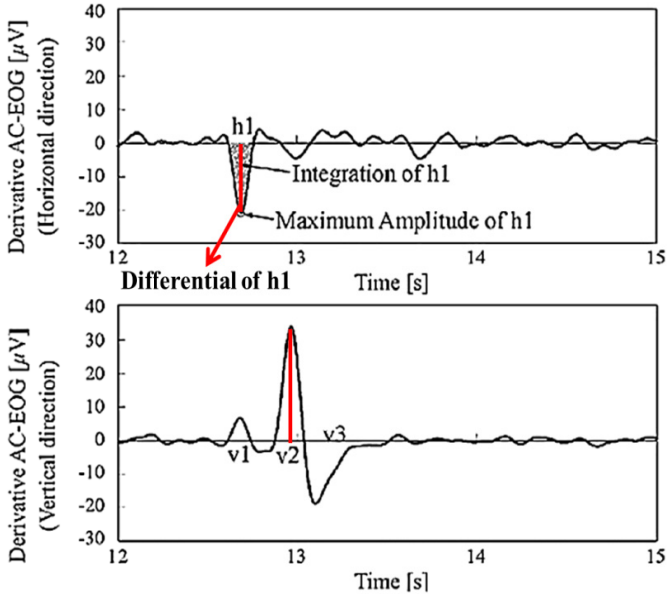


Fig. 5. Example derivative of a horizontal and vertical direction AC-EOG signal

displacement of the preceding movement was calculated as the sum of the displacement values of saccades occurring between the previous and the present eye-gazes.

3.5 Decision Algorithm

Determinations were carried out by focusing on the 0-0 wave at the intersection of the baseline 0V and the differential of the AC-EOG. In our proposed method, the system makes a determination using the following algorithm:

Alg 1.

1. Processing of the high cut-off 20 Hz FIR filter to the AC-EOG begins.
2. Differential processing begins.
3. Processing of the high cut-off 15 Hz FIR filter begins.
4. To reduce steady-state noise by calibration [9].
5. If the threshold value of the integral of the first horizontal 0-0 wave and the second horizontal 0-0 wave have the same sign, reduce the equivalent 2° (If the signs are opposite, take no action).
6. Detect the vertical 0-0 wave in the time series that occurs nearest to the horizontal 0-0 wave.
7. If it is the same as the first eye movement direction determined by the second time, this decision will continue to be discarded.
8. Repeat steps 5, 6, and 7 to determine eye-glance behavior.

4 Result

4.1 Eye-Glance Error Rate during Free Time

As described in the experimental procedure outlined in Section 3.3, two zones of experiment time (eye-glance and free time) were observed. An example of the derivative AC-EOG signals observed during experiment and free time analyzed with the method described in session 3.4 and 3.5 is shown in Fig. 6.

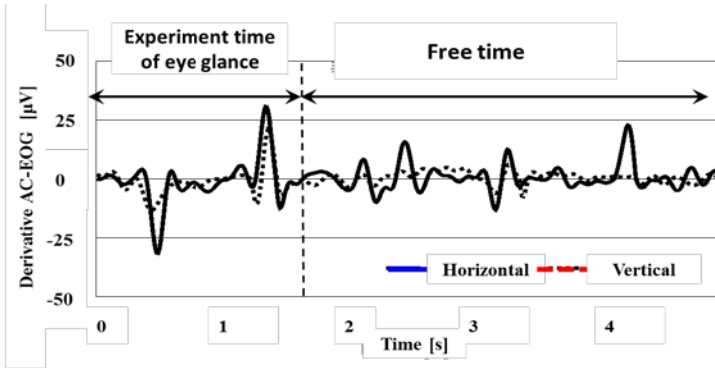


Fig. 6. Derivative AC-EOG signal example showing both experiment and free time

First, we used the characteristics of the horizontal 0-0 wave to determine normal eye movements. We defined error rate that eye-glance have detected during free time. As a result, the error rate average for all subjects during free time was as determined to be 12.73% by the horizontal 0-0 wave, as shown in Table 1.

Table 1. Eye-glance error rate during free time as determined by the horizontal 0-0 wave

Subject	A	B	C	D	Average
Error rate[%]	11.27	14.29	16.44	8.92	12.73

To addition, we determined τ by calibration. The results showed that the average eye-glance error rate for all test subjects during free time was below 5%, as measured by horizontal 0-0 wave, and the pause time τ , as shown in Table 2.

Table 2. Eye-glance error rate during free time by horizontal 0-0 wave and pause time τ

	Pause time range [s]	Subject				Average
		A	B	C	D	
Error rate [%]	$\tau \pm 0.1$	2.81	1.15	1.69	0.32	1.49
	$\tau \pm 0.2$	4.50	2.29	4.06	0.32	2.79
	$\tau \pm 0.3$	6.19	2.29	6.10	0.64	3.81
	$\tau \pm 0.4$	6.75	2.52	7.45	0.96	4.42

4.2 Eye-glance Detection

Using the same method, we attempted to detect eye-glance behavior during experiment times using the horizontal 0-0 wave and the pause time τ . A successful value was counted when an eye-glance was detected at least one once during the experiment time. The results of our experiment showed that the success rates for all test subjects were above 90% as determined by the horizontal 0-0 wave and the time τ , as shown in Table 3.

Table 3. Detection rate of eye-glance by horizontal 0-0 wave and the pause time τ

	Pause time range [s]	Subject				Average
		A	B	C	D	
Detection rate [%]	$\tau \pm 0.1$	43.82	77.54	42.54	60.69	56.15
	$\tau \pm 0.2$	75.14	97.53	75.03	78.57	81.57
	$\tau \pm 0.3$	87.82	97.54	90.41	85.88	90.41
	$\tau \pm 0.4$	87.46	97.45	92.45	85.65	90.75

4.3 Direction Judgment

Offline analyses were performed with the four subjects to compare the new integration method with the method of using three characteristics (as the same time horizontal 0-0 wave and vertical 0-0 wave which the pause time is $\tau \pm 0.3$). The accuracy of a choice as defined as the success rate divided by the number of choices made. Table 4 shows the accuracies of the choices for all subjects along with the sum of all the displacements. In the case of four possible choices using eye-glance behaviors, the mean accuracy of the choices was 79% when the three characteristics were used.

Table 4. The accuracies (%) of choices for all subjects using three characteristics

	Pause time range [s]	Subject				Average
		A	B	C	D	
Detection rate [%]	$\tau \pm 0.3$	81.32	95.04	85.77	55.03	79.29

The direction judgment accuracies for all subjects, including the vertical detection rate, was reduced by 10% compared to measurements of the horizontal direction alone, as shown in Tables 3 and 4. Especially, with respect to the subject D, including the vertical detection rate caused a 35% reduction compared to judgments of the horizontal direction.

5 Discussion

5.1 Timing of Vertical 0-0 Wave

It is possible that the timing of the vertical 0-0 wave included an error, as shown in Fig. 7. The gray zone shows that the timing of the horizontal 0-0 wave, and twice were different, but twice vertical 0-0 wave were the same direction in this example.

One of the reasons was that this subject tended to move his neck as well as his eyes vertical direction when using the smartphone for long periods of time. Furthermore, there are individual differences in the neck movements of all test subjects. In particular, this is why Subject D showed a 35% accuracy reduction.

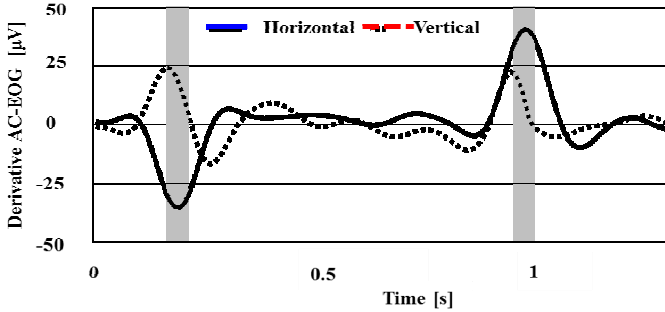


Fig. 7. Direction error judgments for single eye glance behavior

5.2 Character Input with Using Eye-Glance

In an attempt to create a practical application for use with our eye glance input interface; we designed a guide menu for character input that utilizes eye glance behavior and a small screen, as shown in Fig. 7. For our future work, intend to conduct character input experiments using this guide menu.

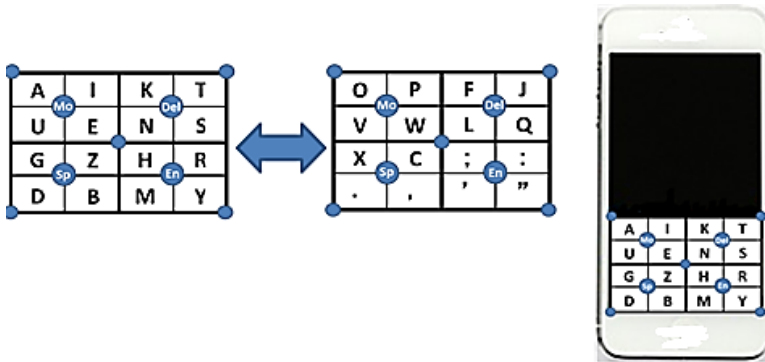


Fig. 8. Guide menu for character input

6 Conclusions

In this paper, we proposed an eye-glance input interface that allows users to make inputs consistently without prior declarations and discussed a decision algorithm for the interface that utilizes a horizontal 0-0 wave and the pause time τ .

During experiments conducted on four test subjects, we achieved a detection rate above 96% for non-glance behavior using the horizontal 0-0 wave and pause time τ .

In contrast, for actual eye glance behavior, the average detection rate of the four subjects was above 90% as measured by the horizontal 0-0 wave and the pause time ($\tau \pm 0.3$).

However, when the vertical detection rate was included, the accuracies of direction judgments for all subjects were 79%, and a 10% reduction occurred when the vertical 0-0 wave was added. Our future tasks and goals include designing an input guide menu for a small screen that can be used with the eye-glance interface along with creating an application suitable for a smartphone or tablet.

References

1. Dekun, G., Naoaki, I., Tota, M., Kazuyuki, M.: Improvement of Eye Gesture Interface System. In: The 16th Asia Pacific Symposium of Intelligent and Evolutionary Systems. Session 3, paper 3-1 (2012)
2. Kazuyuki, I., Yasuo, S., Tooru, I.: Eye Gaze Communication System for People with Severe Physical Disabilities. The Transactions of the Institute of Electronics, Information and Communication Engineers J83-D-I(5), 495–503 (2000)
3. Ryosaku, K., Araki, O., Takashi, N., Nobuyoshi, F., Kei, T., Chiaki, H.: Clinical usefulness in patients with ALS: introduction of the device to communicate in input method gaze. The Journal of Japanese Physical Therapy Association 28, 295 (2001)
4. Hiroshi, S., Yoshiaki, N., Naoki, U., Naoki, H., Hidekazu, Y.: A Prototype of Head-Attached Interface Device and Its Functional Evaluation. The Society of Instrument and Control Engineers 36(11), 972–979 (2000)
5. Hiroshi, S., Yoshiaki, N., Naoki, U., Naoki, H., Hidekazu, Y.: Study of eye-gaze input interface configuration by Eye-Sensing HMD. Human Interface Society 2000, 239–242 (2000)
6. Naoaki, I., Takumi, O., Kazutaka, S.: Investigation for Calculation Method of Eye-Gaze Shift from Electro-Oculograph Amplified by AC Coupling with Using Eye-Gaze Input Interface. The IEICE Transactions on Information and Systems J90-D(10), 2903–2913 (2007)
7. Naoaki, I., Takumi, O., Yutaka, S.: Eye-gaze input interface with head mounted display and electro-oculograph amplified by AC coupling. Human interface. The Transaction of Human Interface Society 9(4), 75–84 (2007)
8. Kazutaka, S., Naoaki, I.: Multi Selection Type Eye-Gaze Input Interface Using Eye Movement of Diagonal Direction with EOG Amplified by AC Coupling. The IEICE Transactions on Information and Systems J92-D(2), 189–198 (2009)
9. Naoaki, I., Kazutaka, S.: A new method for calculating eye movement displacement from AC coupled electro-oculographic signals in head mounted eye-gaze input interfaces. Biomedical Signal Processing and Control 5, 142–146 (2010)
10. Barea, R., Boquete, L., Mazo, M., Lopez, E.: Wheelchair Guidance Strategies Using EOG. Journal of Intelligent and Pobotic Systems 34, 279–299 (2002)
11. Tecce, J., Pok, L., Consiglio, M., O'Neil, J.: Attention impairment in electrooculography control of computer functions. Int. J. Psychophysiol. 55(2), 159–163 (2005)
12. Shiyichiro, K., Ryoko, F., Tatsuo, Y., Nozomi, H.: Method of Menu Selection by Gaze Movement Using AC EOG Signals. The Transactions of the Institute of Electrical Engineers of Japan. C. A publication of Electronics, Information and System Society 129(10), 1822–1827 (2009)

Multi-User Interaction with Shadows

Tomomi Gotoh¹, Takahiro Kida¹, Munehiro Takimoto¹,
and Yasushi Kambayashi²

¹ Department of Information Sciences, Tokyo University of Science, Japan
mune@cs.is.noda.tus.ac.jp

² Department of Computer and Information Engineering,
Nippon Institute of Technology, Japan
yasushi@nit.ac.jp

Abstract. Recent mobile devices such as smart phones exhibit performance as good as desktop PCs, and can be used more intuitively than PCs by using fingers. On the other hand, the defect of such a device is its small size. Its display is just big enough for single user, but is too small for interaction of multi-users. In order to overcome the defect, the research of projecting the display with a handheld projector has expanded. Most of the researches, however, do not allow users to manipulate the projected image in a direct manner. In this paper, we propose operations of projected images through shadows. We can create a shadow by shading the light of the projector with a finger. The shadow can be easily scaled by adjusting the distance between the finger and the projector. Also, since the shadow makes good contrast with the white light of the projector, it can be easily recognized through a camera. Using these properties of the shadow, we have implemented a series of operations required on the desktop, and file transfer as a basic multi-users interaction. We show that the users can perform these operations intuitively with the shadow of two fingertips as if they handle a tablet PC through multi-touches.

1 Introduction

Recently advanced handheld devices such as smart phones or PDAs can be also used as projectors with some special attachments. The capabilities of handheld projectors have also begun attracting attentions as one of wearable devices [1–3]. General usage of such a projector is to magnify the small display of a handheld device. Apart from that usage, the researchers begin to investigate the possibility of such handheld device as a multi-user interaction tool have expanded [4–6]. In the latter case, desktop images projected by two users' handheld devices can be used for not only sharing information but also transferring files through making their icons included in the shared area where the two images overlap. The handheld projectors with cameras make such a technique possible, where each handheld device can recognize the two desktop images and their overlapped area from the same position as the projector through the camera. Once the icon

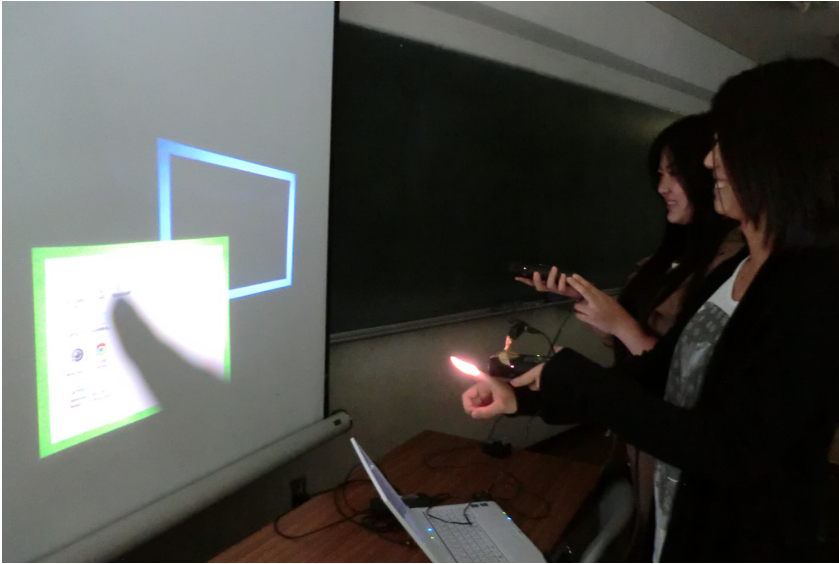


Fig. 1. The projected desktop image with the shadow of a finger

representing a file is recognized, the file can be transferred from one device to the other device, triggered by making the icon move in the overlapped area.

In most cases, these operations are achieved by physically moving the projectors, and therefore thorough operations on the desktop have to be performed at the side of the device. Considering intuitive operations, the direct interaction with projected desktop is preferable rather than operations on the device. It has, however, some problems for practical use. When a user wants to magnify the image projected by his or her handheld projector in order to operate the desktop on a large screen, the user has to look at it from a distance. Since the distance between the projector and the screen is usually longer than human arm, the user cannot touch the projected image directly. Therefore, the user has to utilize a special device such as the laser pointer in order to point a certain object on the screen. Unfortunately, such a special operation is contrary to the intuition augmented by the projector.

In this paper, we propose a set of new desktop operations on projected desktops including multi-user interactions without any special pointing device. Our approach takes advantage of the shadows of fingers on images to operate the desktop instead of real fingers or reflection of laser pointer as shown in Fig. 1. The shadow can be easily created on the projected image by putting a finger between the projector and the screen [7]. Furthermore, the shadow created in this manner can be resized by adjusting the distance of the finger from the projector depending on minuteness of the operation.

Another advantage for using the shadow as an operation tool is that its color and the brightness are relatively stable because the shadow is just a dark area,

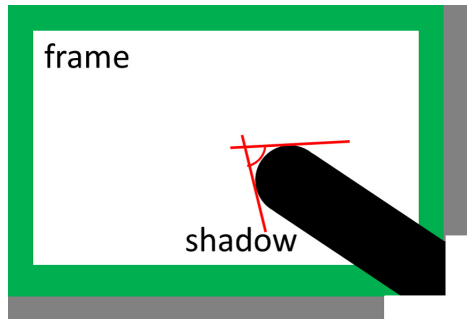


Fig. 2. The shadow of a finger inside a frame

and therefore, it can be easily recognized. Once the black area is recognized, the location pointed by the finger can be also easily detected.

The structure of the balance of this paper is as follows. In the second section, we describe how to recognize the shadows of fingers and fingertips, and then present details of desktop operations using them. In the third section, we show a problem for interactions through the shadow, and then, provide a solution for it. We present the design of our system based on the solution. Finally, we conclude remarks in the fourth section.

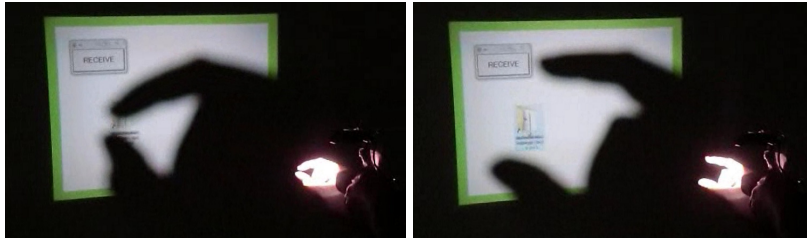
2 Recognition of a Shadow and Operation through It

In this section, we describe how to recognize the shadow of a fingertip, and show how the shadow can be used to operate a desktop. After that, we extend the operation to using two fingers.

2.1 One Finger Operation

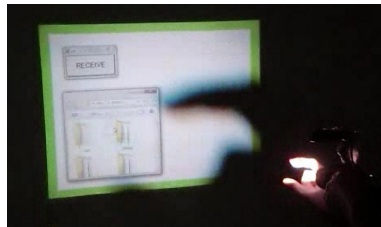
The color and the brightness of a shadow are relatively stable because the shadow is just a dark area, and therefore, it can be easily recognized. Once the black area is recognized and its edge is extracted, the fingertip part of the edge has to be detected. The steps of the detection of the fingertip are follows: 1) draws tangential lines at even intervals, 2) finds neighbor lines such that the angle between them is the smallest, and 3) extracts tangent points between the lines and the finger edge. We determine the location around the tangent points as a fingertip. Fig. 2 shows the smallest angle between tangent lines.

The problem in handling a shadow is that the shadow may be confused with other black objects such as black icons or windows. In order to overcome this problem, we introduce a finger recognition approach based on the frame of a desktop. Notice here that the shadow of a finger has an intersection with the frame. Most other objects on the desktop are inside the frame, and therefore, there is a gap between them and the frame. We identify the black area without



(a) Selection operation with pickup

(b) Open operation with opening fingertips



(c) An opened folder

Fig. 3. Opening a file

such a gap as a finger. As shown in Fig. 2, the shadow created by a finger has an intersection with the frame shown by green border.

For the first step, we have implemented a series of desktop operations on files by a finger: selection, open, drag, and drop. The selection and open are distinguished by the time period in which the shadow of finger stays on the icon. The required time periods are respectively one second and two seconds for the selection and the open, respectively. The drag is made to be active by moving the shadow immediately after the selection. The drop after the drag is performed one second later after the movement of the shadow in the drag stops.

2.2 Two Finger Operation

The one finger operation works well, but since it has to distinguish each operation based on the time in which the shadow stays on a icon, it takes too much time for practical uses, e.g. it takes two seconds for the selection. If the combination of some operations is required, it would take much more time.

In order to reduce the time taking for one finger operation, we introduces a new operation *pickup* that uses two fingers such as a multi-touch on a tablet. In the pickup, we put two fingers together with a small gap such as picking up something. Each operation can be immediately distinguished by pickup immediately

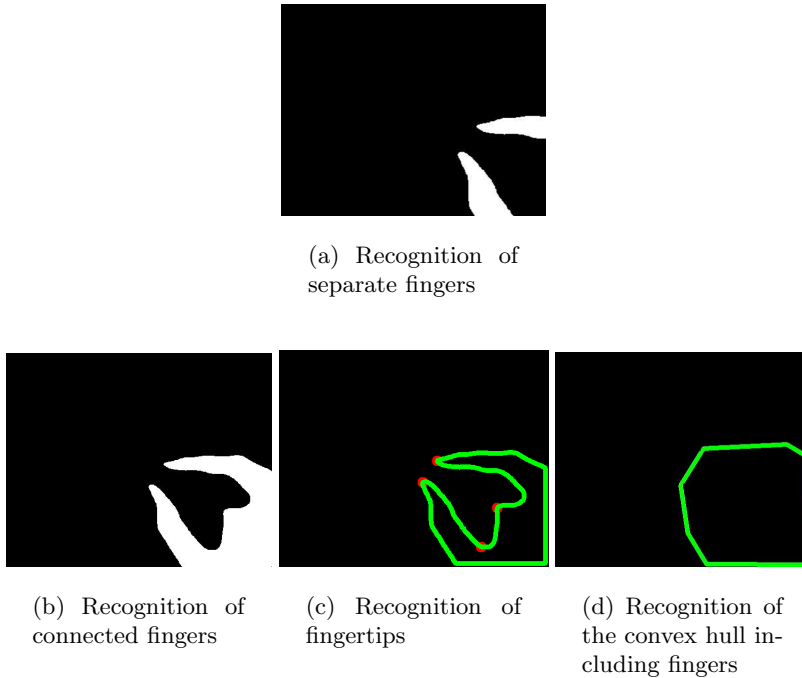


Fig. 4. Recognitions of two fingers

followed by some additional operations. As shown in Fig. 3(a), making pickup held on an icon for a second corresponds to the selection operation. Following it, opening fingers corresponds to the open operation as shown in Fig. 3(b) and (c), or moving fingers holding pickup corresponds to the drag operation.

In order to take advantage of the pickup operation, two fingertips have to be recognized simultaneously. Fig. 2.2 shows the process of recognizing shadows created by two fingers. In Fig. 2.2(a), the shadow looks like two sticks. On the other hand, in Fig. 2.2(b), the shadow looks like a mountain with two tops. We can easily deal with the shadow such as two sticks, because the same recognition technique for one finger can be simultaneously applied to two fingers. On the other hand, it is not easy to recognize the two-top mountain shadow such that shown in Fig. 2.2(c). Red points Fig. 2.2(c) show recognized fingertips in the case of applying one finger recognition technique to the recognized shadow. As shown in the figure, they appear at other than fingertips. That is because the concave angle created by two fingers satisfies the condition of a fingertip.

In order to handle such misrecognized cases, we introduce the technique that identifies the shadow as a convex hull, and ignores recognized points inside the convex hull. For example, the shadow of two fingers shown in Fig. 2.2(b) can be regarded as the convex hull shown in Fig. 2.2(d).

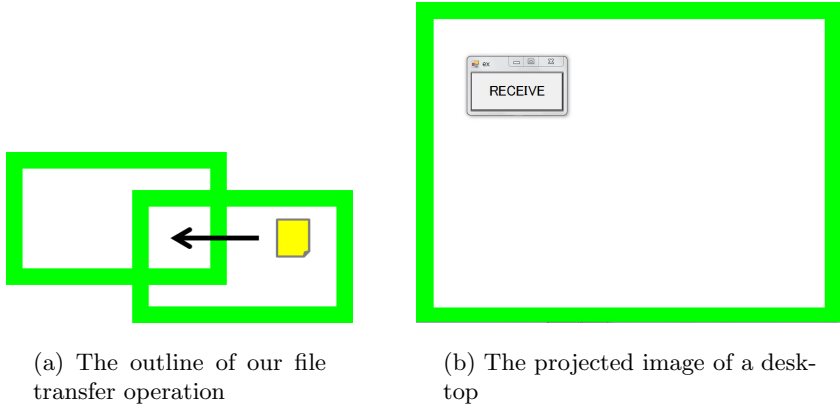


Fig. 5. Interactions of our system

3 File Transfer

The main contribution of using mobile devices through handheld projectors is that it allows several users to share information. Our desktop operation based on shadow allows several users to interact on projected images. Fig. 3(a) shows an example of transferring a file as the basic interaction. In order to transfer the a file, the users overlap projected desktops, which is used as a shared folder, and then the sender drags the icon of the file to the overlapped area i.e. the shared folder. The file transfer manner, which is similar to the operations on tablet PCs, is preferable because the operation is intuitive. It is, however, difficult to make it work based on the shadow. The shadow of a fingertip is created by blocking off the light projected by sender’s projector, and therefore, it disappears in the overlapped area due to the light projected by the receiver.

We mitigate this disappearance problem of the shadow by suppressing the strength of the receiver’s light. Fig. 3(b) shows a desktop image of our system. Each desktop has a receive button inside it, which can be pushed by putting the shadow of a fingertip on it. Once it is pushed, the background of the desktop turns to black, which contributes to suppressing the strength of receiver’s light.

Fig. 3 shows the sequence of the operations for file transfer. As shown in Fig. 3(a), both desktops initially have white backgrounds. Once the receive button on the receiver’s desktop is pushed, the background of the desktop turns to black as shown in Fig. 3(b). After that, overlapping a part of the sender’s desktop on the receiver’s desktop makes receiver’s folder allocated to the overlapped area, as shown by Fig. 3(c). Finally, dragging the icon of the file on the sender’s desktop to the overlapped area, the file is now transferred to receiver. Notice that the shadow of the fingertip also appears on the overlapped area clearly through turning the background, as shown in Fig. 3(d).

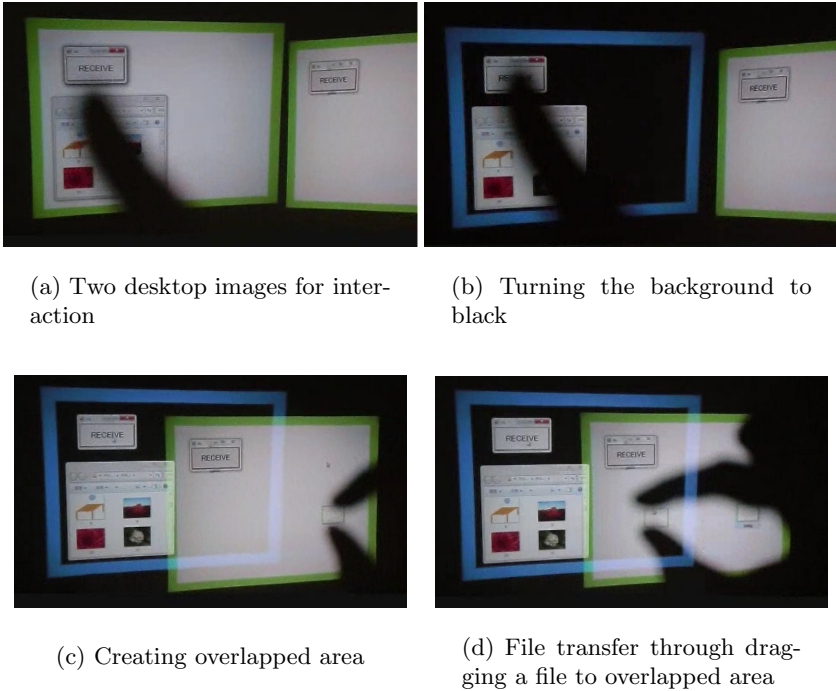


Fig. 6. File transfer

As the last step of the file transfer, the background of the receiver's desktop has to be turned back to white. The background is, however, now black, and the shadow cannot appear on the desktop of the receiver. This means that the button for turning the background back to white cannot be handled by the shadow. Therefore, instead of using shadow, we set the exclusive two kinds of condition where the receive mode is canceled, as follows:

1. The desktops are separated after they are overlapped, or
2. Five seconds passes with keeping the desktops separated.

The first condition naturally realizes the automatic cancellation of the receive mode in a sequence of file transfer operations, and the second condition guarantees that the receive mode is always canceled based on time out.

4 Conclusions

We have implemented a system on which we can operate the desktop using the shadow of fingertips. First, we implemented the operation by using one finger, and then, extended it to the operation by using two fingers. In the one

finger actions, corresponding operations have to be distinguished based on the time period in which the shadow stays on the operated icon, but in the two fingers actions, they can be distinguished based on some actions immediately following picking-up action, which contributes to reducing the total of time cost of operations.

Furthermore, we implemented the file transfer as an example of multi-user interactions based on these actions of the shadows. We designed it as a sequence of operations, in which two desktops are overlapped, and then the operated icon is dragged to the overlapped area. Though it is intuitive, and is a natural extension of multi-user interaction on projected desktop, we observed that it has the problem of disappearance of the shadow in the overlapped area. We then provided a solution for the problem by introducing the technique turning the background of a desktop to black.

The current system is designed for the interaction for two users. We are extending it so that it works for more than two users .

References

1. Cao, X., Balakrishnan, R.: Interacting with dynamically defined information spaces using a handheld projector and a pen. In: Proceedings of the 19th Annual ACM Symposium on User Interface Software and Technology, Montreux, Switzerland, October 15-18, pp. 225–234. ACM (2006)
2. Miyahara, K., Inoue, H., Tsunesada, Y., Sugimoto, M.: Intuitive manipulation techniques for projected displays of mobile devices. In: Extended Abstracts Proceedings of the 2005 Conference on Human Factors in Computing Systems, CHI 2005, Portland, Oregon, USA, April 2-7, pp. 1657–1660. ACM (2005)
3. Yoshida, T., Nii, H., Kawakami, N., Tachi, S.: Twinkle: interface for using handheld projectors to interact with physical surfaces. In: ACM SIGGRAPH 2009 Emerging Technologies. SIGGRAPH 2009, pp. 24:1–24:1. ACM, New York (2009)
4. Cao, X., Forlines, C., Balakrishnan, R.: Multi-user interaction using handheld projectors. In: Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology, Newport, Rhode Island, USA, October 7-10, pp. 43–52. ACM (2007)
5. Mistry, P., Maes, P., Chang, L.: Wuw - wear ur world: a wearable gestural interface. In: Proceedings of the 27th International Conference on Human Factors in Computing Systems, CHI 2009, Extended Abstracts Volume, Boston, MA, USA, April 4-9, pp. 4111–4116 (2009)
6. Yatani, K., Tamura, K., Hiroki, K., Sugimoto, M., Hashizume, H.: Toss-it: Intuitive information transfer techniques for mobile devices using toss and swing actions. *IEICE Transactions* 89-D(1), 150–157 (2006)
7. 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, Newport, Rhode Island, USA, October 7-10, pp. 53–56. ACM (2007)

Intent Capturing through Multimodal Inputs

Weimin Guo, Cheng Cheng, Mingkai Cheng, Yonghan Jiang, and Honglin Tang

School of Computer Science, Beijing Institute of Technology, Beijing 100081 PRC,
Beijing Laboratory of Intelligent Information Technology, Beijing Institute of
Technology, Beijing 100081 PRC
cc@bit.edu.cn

<http://isc.cs.bit.edu.cn/faculties/chengcheng/index.html>

Abstract. Virtual manufacturing environments need complex and accurate 3D human-computer interaction. One main problem of current virtual environments (*VEs*) is the heavy overloads of the users on both cognitive and motor operational aspects. This paper investigated multimodal intent delivery and intent inferring in virtual environments. Eye gazing modality is added into virtual assembly system. Typical intents expressed by dual hands and eye gazing modalities are designed. The reliability and accuracy of eye gazing modality is examined through experiments. The experiments showed that eye gazing and hand multimodal cooperation has a great potential to enhance the naturalness and efficiency of human-computer interaction (*HCI*).

Keywords: Eye tracking, multimodal input, virtual environment, human-computer interaction, virtual assembly, intent.

1 Introduction

Multimodal interaction (*MMI*) is an emerging *HCI* research area which has been developing rapidly in recent years. It is well accommodate the idea of Human-centered design [1]. *MMI* refers to a human-computer interaction paradigm that users utilize a variety of modalities to communicate with computer. Although some natural modalities such as visual input, natural language and gesture input, have their inherent inaccuracy, and is difficult to always meet 100% success rate of recognition, it is sure that they can make the cognitive load of human beings be reduced greatly during interactive process [8]. Employment of visual input modality in a critical *VE* system, e.g. the virtual assembly system, is a first time try to use it to eliminate the ambiguity of user's intents in single-channel input scenarios.

As a natural modality, eye gaze tracking has been developed for human-computer interaction. A number of researchers have been working in this area [7,8,9], and one typical researches is the *Intelligent Gaze-Added Operating System (IGO)* developed by Salvucci and Anderson [10,11]. They compared the performances of the mouse and visual method to show that the visual method has higher error rate than mouse, since the sight often jumps; but the visual input is easier to learn and use, so it is more attractive for users.

Virtual environments are a type of systems established on the base of multimodal interaction. *VE* aim at providing user with an immersive and natural interactive effect, which can hardly be achieved by traditional *WIMP* (Window, Icon, Menu and Pointing) based systems. But the *VE* ability of handling with mixed and fuzzy input data from multiple modalities is still a big problem. Despite researches and developments on hardware have already made a considerable progress and there have emerged a variety of interactive devices, direct manipulation in *VE* today is still a serious technical bottle-neck [4,5]. This point is particularly evident in critical manufacturing environments such as virtual assembly. Building a virtual assembly system with eye tracking modality has an especial merit [6], because eye interaction can free hands from liberous operations.

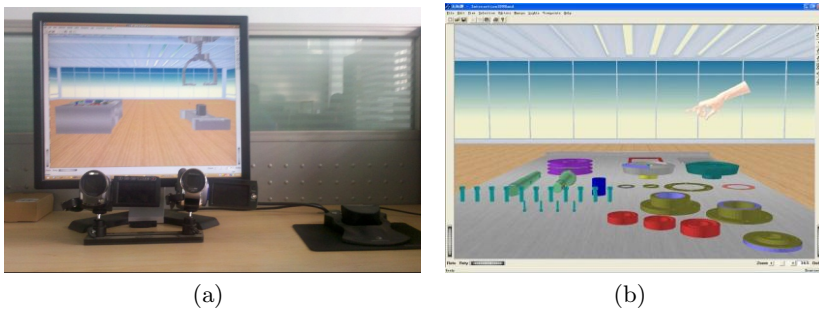


Fig. 1. Multimodal virtual assembly system Interaction3D. (a) System paradigm, (b) Virtual assembly scene.

For this reason we build an eye tracking augmented virtual assembly system *Interaction3D*, as figure 1 shows, where virtual parts are composed with geometric features. One 3D space mouse and a 2D mouse are two modalities to manipulate virtual parts. Two cameras based eye tracking subsystem is another modality to capture user's current focus into the virtual scene. Those input devices work together to constitute multiple inputs. The paper firstly describes eye gaze modality which use *PCCR* technology to estimate user's gaze point on screen [3]. Then introduces how to analyze the gaze point to promote the utility of eye gaze modality in virtual environments.

2 Multimodal Intent Understanding

Gaze tracking provides a potential to allow users naturally express their intents in virtual assembly system. Several main intents are defined in advance in system *Interaction3D*, then multimodal inputs are analyzed to infer users' intents. One of the challenges in gaze tracking is midas touch problem, that is, the randomness of the movement of users' sight [13]. Avoiding this problem is an important part of this research.

2.1 Eye Tracking Modality and Eye Gaze Estimation Algorithm

To accurately catch user's gaze points and from which to understand user's intent is significant. 3D gaze estimation method considers the user's head moving [19]. Relatively, 2D gaze estimation method use a calibrated sight mapping function to estimate the direction of sight [20]. Obviously combining the 2D and 3D gaze estimation methods will be an ideal way. As such we adopt a scheme to achieve this effect, i.e. adapting the map function of 2D gaze estimation to the natural movements of the head by means of head motion compensation.

PCCR (Pupil Center Corneal Reflection) based 2D eye estimation method [3,21] combining a face color and organ model is used to extract the pupil-glint (*PG*) vector. A sampling based *HSI* color space modeling method is adopted here. A real-time liner prediction based eye tracking algorithm is used for extracting pupil-glint vector [24]. Gaze mapping equation obtained from a calibrating process transforms the current user's *PG* vector to the gaze coordinates on view space. The extracted *PG* vector v is denoted as (x_v, y_v) , and the gaze on the screen S_{gaze} is denoted as (x_g, y_g) , the mapping equation $S_g = f(x_v, y_v)$ can be expressed by the following nonlinear equation [2,23], where the parameters a_i and b_i are realized by a 9-point calibrating process :

$$x_g = a_1 + a_2x_v + a_3y_v + a_4x_vy_v + a_5x_v^2 + a_6y_v^2 \quad (1)$$

$$y_g = b_1 + b_2x_v + b_3y_v + b_4x_vy_v + b_5x_v^2 + b_6y_v^2 \quad (2)$$

To eliminate the discrepancy of *PG* vector caused by head movement, an iterative compensation algorithm is used. The main idea of head offset compensation is the conversion function $g(v_2, O_2, O_1)$ which can convert arbitrary *PG* vector to a calibrated position *PG* vector [14]. When we get the corrected *PG* vector, we input it into the gaze mapping function (*EQ1*, *EQ2*) to calculate a new screen gaze point S . This is an iterative process and will converge in less than 5 loops. At last we can obtain a gaze point on the screen [22].

2.2 User Intent Recognition from Multimodal Inputs

Eye gazing can reduce the number of hand modality state switches, so make the interactive process be more smooth. In our virtual assembly environment, one of the most significant intent is feature matching. The so called feature matching refers to that *VE* selecting a feature on target part which has an assembly relationship with a feature on current part. Here we just use this intent to demonstate the capability of multimodal inputs. The scenario of feature matching is like this: user selects one part and assigns it as target part, then selects another part and assigns it as current part, and move the two parts by dual hand respectively. When the current part approaches the target part, a current feature (*curFea*) on current part is determined and a feature (*tarFea*) on target part which can match the *curFea* will be searched by algorithm. For some reasons, there may be several potential candidates of *tarFea* that satisfy the requirements. Eye gaze points a feature to cancel out ambiguity.

Two kinds of eye movement states, fixations and saccades are recognized as basic eye movements [16]. A sequence of original gazing points are clustered by the gazing points clustering algorithm. The gazing points that meet both temporal and spatial constraints are collected into a cluster; Secondly, The center of cluster (*COC*) points are evaluated. Thirdly, determining if the *COC* points fall in a dwelling space threshold of the center of feature (*COF*) points. And at last, some smoothing filter process is done to remove the eye movement noises, as such we extract eye movement behaviors depicted by scan lines.

The collected original signals of viewpoints on screen can be represented as a viewpoint sequence G :

$$G = \{g_i | i = 1, \dots, n\}, g_i = (x_{gi}, y_{gi}, t_{gi}) \quad (3)$$

Gaze point g_i can be denoted by a 2D coordinate (x_{gi}, y_{gi}) on the plane together with the time tag t_{gi} . But the users' gaze focus cannot be expressed directly with these original gazing points G due to the midas touch problem. *COC* points is used to substitute the original gaze data. There are some online clustering algorithm [17], but the known algorithms e.g. successive k-means algorithm, online hierarchical clustering algorithm and general clustering algorithms with unknown number of clustering [18], cannot meet our real-time requirements. Here we give a gaze clustering algorithm concerning both the temporal constraint and spatial constraints. This can be simply explained by the distance metric formula below:

$$d(g_i, g_j) = \frac{k_1 * \sqrt{(x_{gi} - x_{gj})^2 + (y_{gi} - y_{gj})^2} + k_2 * |t_{gi} - t_{gj}|}{k_1 + k_2} \quad (4)$$

Here k_1 and k_2 are weighting coefficients which adjust the importance of spatio-temporal constraints. The viewpoints sequence in the scope of certain spatio-temporal thresholds are chosen to be a viewpoint cluster C_i , and update the current cluster center. The analysis of eye movement is conducted. If the viewpoint continuously falls into a *COF* range, that is, the distance of a *COC* and a *COF* is less than the space threshold, and the duration is more than time threshold, then it can be determined as a visual dwelling, thus it can be thought as focusing on a feature.

Feature matching intent is determined by scenario other than individual eye modality. The eye gaze on interest targets does not always means feature matching intent. Only in a specific perceptive scenario does the eye gaze indicate user current intent in *Interaction3D*. As such, we mainly infer intent with the scenario and eye modality data. Users manipulate virtual parts through dual hand modalities to direct the scenario. Another important role of scenario is *VE*. It is responsible for perceiving spatio-temporal relationships among the virtual objects and gives a representation of specific perception. The virtual objects in the virtual environment, the multimodal inputs and the perceptive representation together compose an interactive scenario. We suppose every modality has

several states. The hand modality state set is { *FreStatic*, *FreTrans*, *FreRot*, *PntGst*, *SwitchTR*, *DisableCN*, *GrspStatic*, *GrspTrans*, *GrspRot*}, etc. For example, *GrspTrans* represents the state that user is grasping and translating a part. The scenario for feature matching intent can be described as: the user grasps and translates two parts with dual hands individually and makes the two parts approach each other. *VE* perceives matched features and highlight the features. User's eye casts a gaze ray through a *tarFea*.

3 Experiments and Discussion

In order to demonstrate the feasibility of eye modality and also explore the way to evaluate and analyze the accuracy of eye modality in *Interaction3D*, we design an eye tracking experiment. The multimodal intent understanding was evaluated in a user study in which computer students were tasked with various feature matching exercises carried out using the system *Interaction3D*.

3.1 Participants

The experiment included 20 participants who are all graduate students in university. They are all volunteers and all agree to be honest to report the experiment results. Because the experiment involves a developing eye tracking system, the participants spend several hours to adapt themselves to the eye gaze tracking subsystem. Before the experiment, the participants were also introduced about the virtual assembly system *Interaction3D*, although participants need not do any real assembling operations.

3.2 Apparatus

The system hardwares primarily are personal computer , two cameras, and a 3D space mouse. The CPU is Intel Core 2,3.0 GHz, with 2G memory. The screen is 19 inch LCD at a resolution of 1280*1024 pixels. The cameras are all *Panasonic HDC-SD60*. Dual cameras are fixed on a bracket and located just below the screen. The video card is *Nvidia Geforce 450*. The 3D mouse is *3Dconnexion Spacemouse XT*.

Software used in the experiment includes two prototype systems, eye tracking system and virtual assembly system , and a small statistical program to experiment itself. *Interaction3D* is developed on virtual environment development platform *Open Inventor 5.0*, and the eye tracking system is on the platform *OpenCV*. During the interactive process, participants eye movement are limited in the range of 10° with an average viewing distance of 50cm. Participants use 3D space mouse to pick and move the parts to a place to garrantee there is a minimal distance 70 pixels between different geomic features. Another software is programmed specifically for the experiment which draws out the eye gaze points with colors indicating timing, gives a statistical results about the clustering of the gaze points, and calculate errors of the gaze points respect to the center of feature.



Fig. 2. Experiment deployment. (a) input devices, (b) multimodal testing.

3.3 Procedure

Participants were seated in front of the computer display at a distance of about 50cm, and were allowed to move heads within a space of approximately 200*200*300 mm (width*height*depth). In the stage of experiment preparation, a process of nine-point calibration is done for every participant. Before the experiment, participants were given a brief practice session to familiarize themselves with the input devices and tasks, see figure 2. During the regular trails, the participants firstly uses 3D space mouse to pick and move a part to the middle of the screen and make it occupying about 1/3 view space, secondly to push an icon on the interface to start eye tracking procedure and gaze at the features one by one for 10 seconds totally. thirdly, the participant push an icon to stop the eye tracking procedure. and the gaze data were collected in a log automatically. Forthly, experiment designer takes a screenshot for each trial. Participants were required to repeat the experiment three times with different part each time. At last, the participants completed an experiment questionnaire to describe which features were gazed at for the different three parts respectively.

3.4 Design

There were five typical parts and totally fifteen features were under consideration. We had named every part and have listed the features of every part in a questionnaire . The participant would just click the parts and number the features to complete the questionnaire.

The gaze point and gaze timing on view plane are drawn with colored dots. Then the gaze points are clustered with red circles, and *COCs* were evaluated and represented with the grey dots, see figure 3.

Evaluation use three metrics, the success rate of feature selection, the error between *COC* and *COF*, and the variance of clusters. When we have got *COC*, we calculate the distance between the points *COCs* and *COFs*, where the latter data comes from the questionnaires. When a distance between a *COC* and a *COF* is less than a threshold d_0 , we say that the feature is successfully selected by the participant.

3.5 Results and Discussion

The experiment results are given in table 1. Three types of feature deployments are separately tested. The corresponding explicit demonstrations of the cases are shown in figure 3. We can discover from the table that the feature selection success rates are very high for all the three types. The average errors between *COCs* and *COFs* were less than 10 mm. If the error threshold was 15mm, the features that had reported by the participants would nearly all be successfully selected. We can see from the data that the error for linear feature deployment is a little bit less than these for the other two deployments. The average error variances of the three types are all big values, and the values for array deployment and radial deployment are bigger than that of the linear deployment.

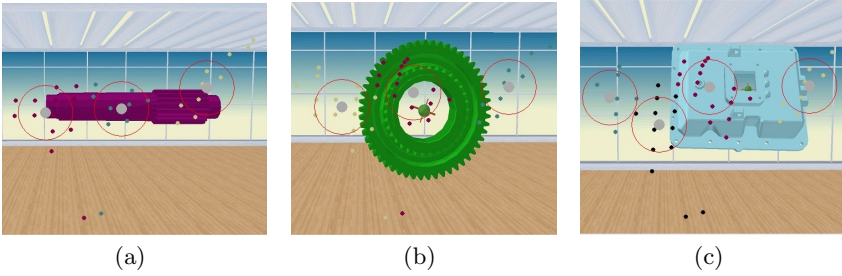


Fig. 3. Experiment examples. (a) shaft features, (b) gear features, (c) casecap features.

The errors between *COCs* and *COFs* come from the eye tracking algorithms used in the eye modality. It makes the user's sight have certain uniform deviation and it is less accurate in vertical direction. This measure can be considered as system error that should be minimized. The variances of errors represent the source gaze points distribution. The results illustrate that the estimated eye gaze points themselves can not be directly used in intent inferring, while *COC* substitutes the source data makes eye modality more practical. This measure can be considered as a human factor that can be reduced. We can discover from the experiment result source data that the variances vary between the different participants. This confirms the statement that human's eyes are continuously moving when they look at something [25].

Table 1. The experiment result

FeatureDeploy	SelSuccessRate	AveFocusError	ErrorVariance
Linear	98.4%	7.26(mm)	79.21
Array	96.7%	9.86(mm)	93.36
Radial	97.1%	9.53(mm)	81.29

The experiment design considered only the eye gaze feature selection in static scene. This is because analysis of eye modality accuracy is very critical for multimodal intent inferring. From the static scene gaze evaluation, we really discovered the factors that make effects to the eye modality and have found the way to promote the efficiency of eye modality. Because user direct manipulations in virtual environments are very slow compared with eye tracking frame rate, the gaze points cluster analysis in dynamic assembly operations has not much difference from that of static situation. The experiment results give us confidence to using eye modality to infer user's intents.

4 Conclusion

In this paper we presented a novel research on intent capturing in virtual assembly environments. We built a prototype of multimodal virtual assembly system and use eye gaze to help system infer feature matching intent. Although only one intent is analyzed in this paper, the work are fundamental for future intent driven virtual environment system construction. With respect to virtual assembly, eye modality can be used to substantially reduce the cognitive overload needed for designers by inferring their interactive intents. An experiment was designed to validate the feasibility of eye modality and the potential of using multimodal input to infer user's intents in virtual environments. Results of the experiments were favorable. The online eye gaze cluster algorithm can provide an stable and accurate feature selection in virtual assembly scenario. This means that it can well support *VE* system to choose a feature from a set of candidates and as such infer the feature matching intent in the scenario.

The work introduced in this paper is still preliminary. We have just finished a specific intent capturing with eye gaze modality. Next we will explore more intents with eye gaze modality. In the following research we need to continue to study how to make eye gaze modality more applicable to designers. We will design an experiment which can validate the eye gaze modality in a dynamic multimodal situation.

References

1. Bolt, R.A.: The Human Interface. Lifetime Learning Press, California (1984)
2. Argue, R., Boardman, M., Doyle, J., Hickey, G.: Building a Low-Cost to Track Eye Movement. Faculty of Computer Science, Dalhousie University (December 9, 2004)
3. Guestrin, D., Eizenman, M.: General Theory of Remote Gaze Estimation Using the Pupil Center and Corneal Reflections. *IEEE Transaction on Biomedical Engineering* 53(6), 1124–1133 (2006)
4. Chen, M., Luo, J., Dong, S.: Task-Oriented Synergistic Multimodality. In: Proceedings of the First Interactional Conference on Multimodal Interface(ICMI 1996), pp. 30–33. Tsinghua University Press, Beijing (1996)
5. Lin, Y., Chen, M., Luo, J., et al.: An Architecture for Multimodal Multi-Agent Interactive System. In: Proceedings for Interactional Conference on CAD/CG 1997. Interactional Academic Press, Beijing (1997)

6. Ford, N.: Cognitive Styles and Virtual Environments. *Journal of the American Society for Information Science* (April 2000)
7. Dong, S.H., Chen, M., Luo, J.: The Model, Method and Instance of Multimodal User Interface. *Universitatis Pekinences* 34(2-3) (April 1998)
8. Fang, Z.G.: Visual Line Tracking Technology and Its Application in Multimodal User Interface. *Systems Engineering and Electronics*
9. Jacob, R.J.K.: What You Look at is What You Get: Eye Movement-Based Interaction Techniques. In: *CHI 1990 Poceeding*. ACM (1990)
10. Salvucci, D.D., Anderson, J.R.: Intelligent gaze-added ingerface. In: *CHI 2000 Conference Proceedings*. ACM Press, Netherlands (2000)
11. Kumar, M., Paepche, A., Winograd, T.: EyePoint Practical Pointing and Selection Using Gaze and Keyboard. In: *CHI 2007, San Jose, CA, USA*, pp. 421–430. ACM Press (2007)
12. Tanriverdi, V., Jacob, J.K.: Interaction with eye movements in virtual environments. In: *CHI 2000 Conference Proceedings*. ACM Press, Netherlands (2000)
13. Barfield, W., Furness, T.A.: *Virtual Environments and Advanced Interface Design*. Oxford University Press, Oxford (1995)
14. Zhu, Z., Ji, Q.: Eye Gaze Tracking Under Natural Head Movements. In: *Proceeding of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR* (2005)
15. Kumar, M., Klingner, J., Puranik, R., Winograd, T., Paepcke, A.: Improving the Accuracy of Gaze Input for Interaction. In: *ETRA 2008 - Proceedings of the Eye Tracking Research and Application Symposium*, vol. 26-28, pp. 65–68 (March 2008)
16. Carpenter, R.H.S.: *Movements of the Eyes*. Pion Ltd. (1988)
17. Omran, M.G.H., Engelbrecht, A.P., Salman, A.: An Overview of Clustering Methods. *Intelligent Data Analysis* 11(6), 538–605 (2007)
18. Guha, S., Meyerson, A., Mishra, N., Motwani, R., O' Callaghan, L.: Clustering data streams: Theory and Practice. *IEEE Transaction on Knowledge and Data Enginerring* 15(3), 515–528 (2003)
19. Beymer, D., Flickner, M.: Eye gaze tracking using an active stereo head. In: *Proceedings of the Computer Vision and Pattern Recognition, Madison*, pp. 451–458 (2003)
20. Morimoto, H., Mimica, M.: Eye gaze tracking techniques for interactive applications. *Computer Vision Image Understand, Special Issue on Eye Detection and Tracking* 98, 4–24 (2005)
21. Matsumoto, Y., Zelinsky, A.: An algorithm for real-time stereo vision implementation of head pose and gaze direction measurement. In: *Proc. 4th IEEE Int. Conf. Automatic Face & Gesture Recognition*, pp. 499–504 (2000)
22. Shih, S.W., Liu, J.: A Novel approach to 3-d Gaze Tracking using Stereo Cameras. *IEEE Trans. Syst., Man and Cybern., Part B*, No 1, 234–245 (2004)
23. Duchowski, A.T.: *Eye Tracking Methodology: Theory and Practice*. Springer (2002)
24. Cheng, C., Jingjing, D.: Research and Realization on Real-time Linear Prediction Algorithm for Desktop Eye-Tracking. *Acta Electronic Sinica* 37(4A) (April 2009)
25. Slater, M., Steed, A., Chrysanthou, Y.: *Computer Graphics and Virtual Environments: From Realism to Real-time*. Addison Wesley (2002)

Robust Hand Tracking in Realtime Using a Single Head-Mounted RGB Camera

Jan Hendrik Hammer¹ and Jürgen Beyerer^{1,2}

¹ Karlsruhe Institute of Technology (KIT)
jan.hammer@kit.edu

² Fraunhofer Institute of Optronics, System Technologies and Image Exploitation,
Karlsruhe, Germany
juergen.beyerer@iosb.fraunhofer.de

Abstract. In this paper novel 2D-hand tracking algorithms used in a system for hand gesture interaction are presented. New types of head-mounted Augmented-Reality devices offer the possibility to visualize digital content in the user's field of view. To interact with these head-mounted devices hand gestures are an intuitive modality. Generally, the recognition of hand gestures consists of two main steps: The first one is hand tracking and the second step gesture recognition. This paper concentrates on the first step: Hand tracking. Due to the wearing comfort of the glasses-like systems these only use a single camera to capture the field of view of the user. Therefore new algorithms for hand tracking without depth data are presented and compared to state-of-the-art algorithms by utilizing a thorough evaluation methodology for comparing trajectories.

1 Introduction

The development of mobile glasses-like Augmented-Reality (AR) devices is striding along. Different companies are working on these so called *high-tech glasses* or *cyberglasses*. Besides the capability of offering optical see-through AR the only head-mounted device (HMD) having eye tracking functionality is the *Interactive See-through HMD*¹. The HMD has further been extended with a scene camera for capturing the user's field of view and serves as core device of the European project *ARTSENSE*². The goal of *ARTSENSE* is to develop a system enhancing the experience of a museum visit by providing the visitor with digital content adapted to his personal interest [5]. Gaze is used to implicitly detect the visual attention [17] that heavily contributes to the decision of what is of interest to the visitor. Hand gesture recognition is used to detect intuitive and easy to learn wiping- and pointing-gestures making explicit interaction with the

¹ <http://www.interactive-see-through-hmd.de/>

² Augmented Reality Supported adaptive and personalized Experience in a museum based on processing real-time Sensor Events, funded by the European Commission under the 7th Framework Program, Grant Agreement Number 270318.

system and visual AR content possible. Basis of a gesture recognition is the hand tracking that we will focus on in this paper. Since depth sensors are too heavy to be attached to an HMD, we concentrate our work on 2D-hand tracking with a single RGB camera. The structure of this document is as follows: In Sec. 2 we present related work. Afterwards, in Sec. 3 we detail our developed methods and in Sec. 4 the used evaluation methodology. Before the conclusion and outlook in Secs. 6 we give information on the real-time capability in Sec. 4.

2 Related Work

In mobile applications the following challenges are prevalent: Lighting conditions may change constantly and – since the sensor is head-mounted – the background is not static. Both make the process of hand localization more difficult as in stationary applications, where simple frame differencing yields high quality hand segmentations [2] or trustable motion information [18]. That is why in mobile applications researchers use gloves [21], markers [11], accelerometers [16] or thermal cameras [1]. 3D-sensors are widely and successfully used for hand tracking [13]. The problem is that all these approaches are not appropriate for the given scenario. Nothing shall be attached to the hands of a museum visitor and the head-mounted device shall be lightweight in order to be as non-intrusive as possible. Pisharady et al. [15] detect hand postures against complex backgrounds, but their method is far away from being real-time capable. An application similar to ours is that of Kölsch and Turk [9].

3 Hand Tracking

In this section the hand tracking algorithms evaluated in this paper are described. Using only one visual camera we take into account two cues – as most recent procedures do: Skin color and motion information [20]. A rough overview of our system is depicted in Fig. 1. The input images are fed into a segmentation process which deals with the localization of the hand using skin color detection and motion information. Afterwards, the hand is tracked using different approaches including particle filters or Flocks of Features [9]. As result the tracking has a trajectory used e.g. for gesture recognition. In Sec. 3.1 we describe our approach for skin color detection and in Sec. 3.2 for motion computation. Then we detail our hand tracking methods in Sects. 3.3, 3.4 and 3.5.

3.1 Skin Color Detection

Skin color detection is well known especially by research topics like face detection from the last decades. The first question is the one for the color space to be used, the second one for the model representing the skin color distribution. For this purpose parametric models like single Gaussian distributions, Gaussian mixture models (GMM) or non-parametric histograms have been tested.

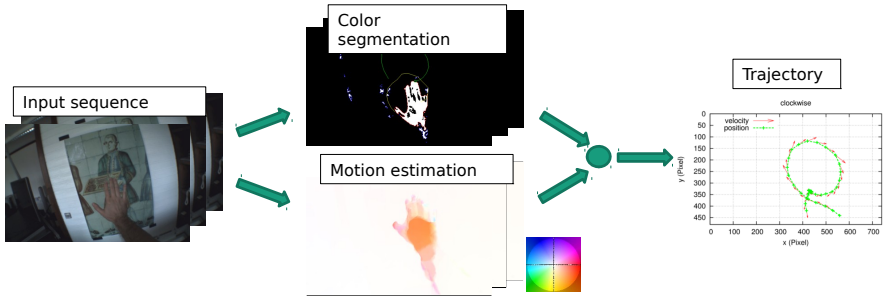


Fig. 1. General overview of our hand tracking algorithms and the used cues

Regarding the color space it has been shown in our experiments as well as the literature ([8], [14]) that 3D color spaces are the ones to choose compared to 2D color spaces. We decided to use RGB since there exists no preference whether RGB or HSV is more suitable.

Histograms do not make an assumption about the distribution of the color to be tracked. Gaussian distributions only work well if the tracked color is distributed normally. GMMs can adapt to not-normally distributed color distributions better, but the number of Gaussian distributions and the weighting factors have to be estimated using an Expectation Maximization step. In comparison, it can be said, histograms can adapt better to special color distributions, but their generalization capability is not as high as that of parametric models.

Independent of the type of model chosen for color representation, training data is needed to fill a histogram or compute the parameters of a Gaussian distribution. In some databases 1000 of images have been annotated according to “what is skin” and “what is not skin”. Furthermore, GMMs have been generated out of this data. Models trained like this show a high generalization capability but tend to be not accurate enough in special situations [7]. We can confirm that circumstance for our case.

A Bayesian classifier along with a threshold [14] is used to produce a binary mask as seen in Fig. 1. For training of the models we use an image patch of a skin-colored region, which can be determined in a calibration phase. The binary mask is afterwards processed with morphological operations to reduce the noise of the segmentation.

3.2 Robust Motion Information

With only one camera and an inhomogeneous background it is not possible to perfectly segment the hands based on color information [1]. By using motion information it is possible to distinguish between different objects in the scene. Motion information is mostly produced by simple frame differencing in stationary applications [18]. Since this does not work for mobile applications, Koelsch and Turk [9] used KLT-features [10] to compute the optical flow for their Flocks of

Features tracker. We estimate motion by computing the optical flow between two images using the FlowLib [22] producing much more precise optical flow [3]. It contains the movement of each pixel to its position in the next frame. The right picture of Fig. 2 displays, how optical flow can be color-coded [3]. The direction of a motion vector is determined by the hue and its length by the saturation. As it can be seen in Fig. 2, the hand moves almost vertically upwards. At the same time, that part of the arm, which is at the lower right margin of the image, moves more to the right. One of the biggest problems concerning optical flow is that algorithms computing accurate flow fields in realtime are rarely available. In Secs. 3.4 and 3.5 is described how motion information is utilized for tracking.



Fig. 2. Left: Hand with overlaid motion vectors. Middle: Color-coded dense flow field. Right: Color wheel for color-coding flow vectors [3].

3.3 Region-Based Hand Tracking

Region-based hand tracking is a simple algorithm relying on skin color detection as described in Sec. 3.1. At first the biggest area of contiguous skin pixels is determined. The *center*-tracking approach only determines the mass value of this biggest skin-colored blob as visualized by the green dot in Fig. 3 on the left. The problem with *center*-tracking can already be seen in that image: If the arm is skin-colored, the hand position determined is distracted from the center of the back of the hand, so tracking becomes inaccurate or fails. To solve this problem we developed *tip*-tracking shown in Fig. 3 on the right. First, the pixel of the skin-colored blob with the smallest y-coordinate is determined. This smallest y-coordinate will be the y-coordinate of the final hand localization. Then, a special hand height corresponding to the dimensions of the hand in the image is chosen determining the height of the green bar shown in the image. The x-coordinate of the hand localization is then computed as the mass value of all blob-pixels under the green bar. This localization is done in each frame resulting in the trajectory of the hand.

3.4 Hand Tracking Using a Particle Filter

A particle filter [6] is a stochastic tracking algorithm. Three things have to be defined: the state, the motion model and the observation model. The simplest

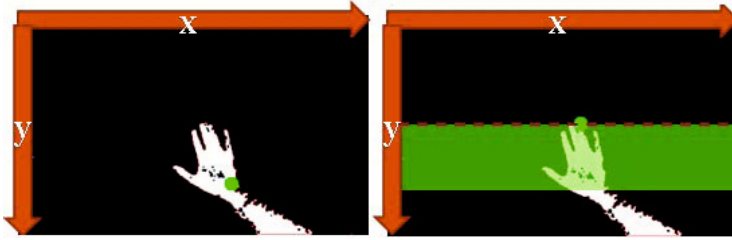


Fig. 3. Left: *center-tracking*. Right: *tip-tracking*.

state of a particle we tested contains one position and one velocity. Often only the skin color probability of the pixel at the particle's position is utilized as observation model, but we found out that using the sum of skin-colored pixels in a local square neighborhood leads to a much higher tracking robustness of the particle filter. This is called the *window-observation* model. As for the *center-tracking* of the previous section, the hand again can be lost because this observation model also computes high weights respectively quality for particles being on the skin-colored arm. Therefore we developed the *shape-observation* model, which is visualized in Fig. 4. The number of skin-colored pixels occluded by the green inner half circle increases and occluded by the yellow outer half circle decreases a particle's weight. Accordingly, the particle on the left in Fig. 4 has a higher weight than the particle on the right. Using this *shape-observation* model particles are prevented from staying on a skin-colored arm. The actual

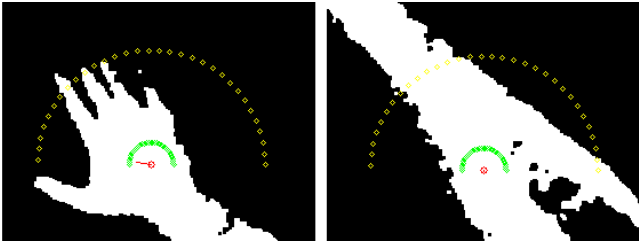


Fig. 4. Left: *shape-particle* with high weight. Right: *shape-particle* with low weight.

velocity as part of the state is determined by the difference vector of the actual and previous position. This motion model is below called *std-motion* model. To further improve the motion of particles we make use of the displacement vectors computed by the FlowLib (cf. Sec. 3.2). This is called the *flow-motion* model.

3.5 Adapted Flocks of Features Tracker

Flocks of Features tracking [9] is one of the state-of-the-art algorithms for hand tracking. This algorithm uses a Viola-and-Jones like detector [19] for the first

localization of the hand. Tracking goes on by using skin color and motion information. Therefore features are placed on the initially found location of the hand and are coupled in a flock by using specific conditions imposed on the feature positions [9]. Motion information is estimated by using KLT-features [10] and skin color probabilities are only considered at the corresponding feature locations.

The Flocks of Features algorithm has been adapted as follows: First, the weight computation of the features has been changed. Originally this computation only considers the skin color probability at the feature's position (*point-mode*). Our version uses all skin color probabilities in a local square neighborhood (*window-mode*) similar to the *window-observation* model of the particle filter described above. Second, instead of KLT-features we use optical flow estimated by the FlowLib [22].

4 Evaluation of Hand Tracking

No common benchmark for hand tracker comparison exists. Because of that we recorded several wiping gestures under different lighting conditions and implemented an evaluation methodology based on the metrics for trajectory comparison found in Needham and Boyle [12]. In Sec. 4.1 the evaluation methodology is described and in Sec. 4.2 the evaluation results are summarized.

4.1 Evaluation Methodology

Our evaluation methodology is based on the metrics for trajectory comparison of Needham and Boyle [12]. Using these makes a thorough evaluation of tracking results possible, since not only detection rates, like the hit rate, false alarm rate or precision of the detection results can be compared. Statistical measures as the median or mean of the deviations between two trajectories allow for precise conclusions. The ground truth trajectories were labeled manually. When labeling is repeated several times, the same person produces similar but slightly different trajectories for the same sequence. Therefore we computed the average distance of manually labeled ground truth trajectories of the same video. The result is an average distance of around five pixels between such trajectories. If an algorithm produces this result, the tracking can be considered as very accurate.

If two different tracking algorithms are compared, it has to be regarded that some track the center of the back of the hand and some the most upper skin pixel (cf. Sec. 3.3). The resulting trajectories are almost the same but shifted by a constant displacement. Hence, one must compensate for this offset by shifting one of the sequences according to this average displacement. After that, scores like the average distance become meaningful. The same yields for the recognition of unreferenced gestures, where not the exact positions are required but the relative trajectories. In this case, since a trajectory can also be seen as sequence of velocities or displacement vectors, the absolute difference of corresponding velocities of compared trajectories should be as small as possible. In the evaluation presented below we consider these velocity deviations as quality metric.

Our benchmark consist of four videos recorded at 25 fps and a resolution of 752×480 pixels. The videos were recorded under dark and light lighting conditions, in front of a complex background. Each contains sixteen wiping-gestures performed by one person. The sequences were recorded for two persons performing gestures with different speeds resulting in more than 5600 frames in total with one hand visible in approximately 50% of the time. The tracking methods were evaluated on these sequences of gestures, in which the hand is entering the field of view of the camera, performing the gesture and leaving the field of view for every gesture. This adds additional difficulty because this entering and leaving must be detected correctly.

4.2 Evaluation Results

In this section we present a comparison of the algorithms described above. All of these show good tracking results on single gestures, but some fail on sequences of gestures. Our evaluation of the Flocks of Features variants revealed improved tracking robustness when using *window*-mode and FlowLib-motion. Still, our best Flocks of Features implementation sometimes fails in detecting the hand leaving the image. As consequence the method starts tracking the background. The adjustment of the Camshift algorithm for properly detecting hands leaving the image is difficult, but Camshift handles this difficulty much better and produces only a few false positives. Camshift further has to be carefully adapted to the hand size and image resolution, which is the same for the *tip*-tracking. Additionally, both solely rely on skin color detection. This is their biggest disadvantage, because if skin color detection fails, they fail tracking. Below we present the results on one of videos containing a sequence of gestures. The following methods have been compared:

1. Region-based hand localization with *tip*-tracking (cf. Sec. 3.3)
2. Particle filtering with 500 and 5000 particles (cf. Sec. 3.4):
 - (a) *std*-motion and *shape*-observation model
 - (b) *flow*-motion and *window*-observation model
3. Camshift [4]
4. Flocks of Features tracking with *window*-observation mode and FlowLib-motion (cf. Sec. 3.5)

In Fig. 5 we see the visualized velocity deviations. It can be seen that all methods can handle this long video with good accuracy. The region-based method *tip*-tracking works very good with a hit rate of above 95% and a false alarm rate of below 2%. The Camshift algorithm produces also good accuracy but has slightly worse hit rates of below 90%. The particle filter variants produce hit rates of above 90% and false alarm rates below 2%. Generally, it has to be underlined that the *window*-observation mode in combination with the *std*-motion model fails tracking but in combination with the *flow*-motion tracking succeeds. The *shape*-observation model even shows similar accuracy without using optical flow. The particle filter variants with 5000 particles produce more accurate results

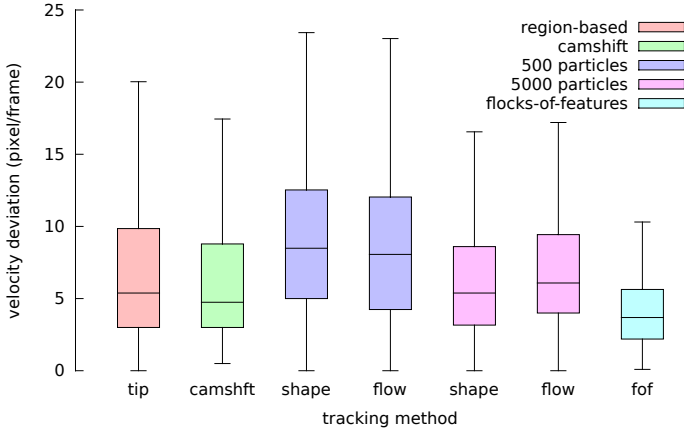


Fig. 5. Comparison of best tracking methods on all gestures with dark lighting conditions

than with 500 particles. The Flocks of Features variant shows the best accuracy on this sequence of gestures but as already mentioned above fails tracking on others videos containing sequences of gestures.

To conclude the evaluation, the developed algorithms have been extensively tested and compared on real sequences using the described evaluation methodology. The results on one sequence of gestures have been presented, which mainly reflect the trackers' overall performance. We could show that our novel observation models incorporated into the particle filter and our adaption of the Flocks of Features tracker as well as the usage of motion information of much higher quality, result in higher tracking accuracy and less tracking failure. However, future tests on other videos with more challenging lighting conditions have to be conducted, because changing lighting conditions directly affect the skin color segmentation. Therefore adaptive skin color models have to be utilized and implemented in all approaches.

5 Realtime Capability

Since the hand tracker is used in an interactive system, it must be realtime capable. *Tip*-tracking reaches 143 fps and Camshift 130 fps. In Sec. 3.2 we mentioned already that robust optical flow unfortunately has a high computational load. The FlowLib in its standard configuration is not realtime capable on the tested image resolution even on modern graphics devices with more than 1000 cores. The *shape*-variant of the particle filter using 500 particles runs with 95 fps. Using 5000 particles reduces the frame rate to 18 fps. The Flocks of Features variant using KLT-features and *window*-observation mode runs with 55 fps.

6 Conclusion and Outlook

To sum up, we have shown new 2D-hand tracking algorithms with increased tracking accuracy and robustness against tracking failure compared to standard approaches as Flocks of Features, Camshift or standard particle filtering. This was demonstrated by realizing a benchmark and an evaluation methodology with different metrics for a thorough trajectory comparison. In the future our focus lies on adaptive skin color models to be able to handle varying lighting conditions and the estimation and fusion of high quality motion information in realtime. To this purpose the benchmark is going to be extended with additional videos containing different people performing gestures in front of different backgrounds under various lighting conditions.

References

1. Appenrodt, J., Al-Hamadi, A., Elmezain, M., Michaelis, B.: Data gathering for gesture recognition systems based on mono color-, stereo color- and thermal cameras. In: Lee, Y.-h., Kim, T.-h., Fang, W.-c., Ślęzak, D. (eds.) FGIT 2009. LNCS, vol. 5899, pp. 78–86. Springer, Heidelberg (2009)
2. Bader, T., Räßle, R., Beyerer, J.: Fast invariant contour-based classification of hand symbols for hci. In: Jiang, X., Petkov, N. (eds.) CAIP 2009. LNCS, vol. 5702, pp. 689–696. Springer, Heidelberg (2009)
3. Baker, S., Scharstein, D., Lewis, J., Roth, S., Black, M., Szeliski, R.: A database and evaluation methodology for optical flow. *International Journal of Computer Vision* 92, 1–31 (2011)
4. Bradski, G.R.: Real time face and object tracking as a component of a perceptual user interface. In: *Proceedings of the 4th IEEE Workshop on Applications of Computer Vision (WACV 1998)*. IEEE Computer Society, Washington, DC (1998)
5. Damala, A., Stojanovic, N., Schuchert, T., Moragues, J., Cabrera, A., Gilleade, K.: Adaptive augmented reality for cultural heritage: Artsense project. In: Ioannides, M., Fritsch, D., Leissner, J., Davies, R., Remondino, F., Caffo, R. (eds.) *EuroMed 2012*. LNCS, vol. 7616, pp. 746–755. Springer, Heidelberg (2012)
6. Isard, M., Blake, A.: Condensationconditional density propagation for visual tracking. *International Journal of Computer Vision* 29, 5–28 (1998)
7. Jones, M.J., Rehg, J.M.: Statistical color models with application to skin detection. *International Journal of Computer Vision*, 274–280 (1999)
8. Kakumanu, P., Makrogiannis, S., Bourbakis, N.: A survey of skin-color modeling and detection methods. *Pattern Recognition* 40(3), 1106–1122 (2007)
9. Kölsch, M., Turk, M.: Fast 2d hand tracking with flocks of features and multi-cue integration. In: *CVPRW 2004 Conference on Computer Vision and Pattern Recognition Workshop*, p. 158 (June 2004)
10. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: *Proceedings of the 7th International Joint Conference on Artificial Intelligence, IJCAI 1981*, vol. 2, pp. 674–679. Morgan Kaufmann Publishers Inc., San Francisco (1981)
11. Mistry, P., Maes, P.: Sixthsense: a wearable gestural interface. In: *ACM SIGGRAPH ASIA 2009 Sketches*, pp. 11:1–11:1. ACM, New York (2009)

12. Needham, C.J., Boyle, R.D.: Performance evaluation metrics and statistics for positional tracker evaluation. In: Crowley, J.L., Piater, J.H., Vincze, M., Paletta, L. (eds.) ICVS 2003. LNCS, vol. 2626, pp. 278–289. Springer, Heidelberg (2003)
13. Oikonomidis, I.: Tracking the articulated motion of two strongly interacting hands. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2012, pp. 1862–1869. IEEE Computer Society, Washington, DC (2012)
14. Phung, S., Bouzerdoum, A.S., Chai, D.S.: Skin segmentation using color pixel classification: analysis and comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27(1), 148–154 (2005)
15. Pisharady, P., Vadakkepat, P., Loh, A.: Attention based detection and recognition of hand postures against complex backgrounds. *International Journal of Computer Vision* 101, 403–419 (2013)
16. Prisacariu, V., Reid, I.: Robust 3d hand tracking for human computer interaction. In: 2011 IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG 2011), pp. 368–375 (March 2011)
17. Schuchert, T., Voth, S., Baumgarten, J.: Sensing visual attention using an interactive bidirectional hmd. In: Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction, Gaze-In 2012, pp. 16:1–16:3. ACM, New York (2012)
18. Spruyt, V., Ledda, A., Geerts, S.: Real-time multi-colourspace hand segmentation. In: 2010 17th IEEE International Conference on Image Processing (ICIP), pp. 3117–3120 (September 2010)
19. Viola, P., Jones, M.: Robust real-time object detection. *International Journal of Computer Vision* (2001)
20. Wachs, J.P., Kölsch, M., Stern, H., Edan, Y.: Vision-based hand-gesture applications. *Commun. ACM* 54, 60–71 (2011)
21. Wang, R.Y., Popović, J.: Real-time hand-tracking with a color glove. *ACM Trans. Graph.* 63, 1–63 (2009)
22. Werlberger, M., Pock, T., Bischof, H.: Motion estimation with non-local total variation regularization. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA (June 2010)

Multimodal Feedback in First Encounter Interactions

Kristiina Jokinen

University of Helsinki, Finland
kristiina.jokinen@helsinki.fi

Abstract. Human interactions are predominantly conducted via verbal communication which allows presentation of sophisticated propositional content. However, much of the interpretation of the utterances and the speaker's attitudes are conveyed using multimodal cues such as facial expressions, hand gestures, head movements and body posture. This paper reports some observations on multimodal communication and feedback giving activity in first encounter interactions, and discusses how head, hand, and body movements are used in conversational interactions as means of *visual interaction management*, i.e. unobtrusive ways to control the interaction and construct shared understanding among the interlocutors. The observations and results contribute to the models for coordinating communication in human-human conversations as well as in interactions between humans and intelligent situated agents.

Keywords: multimodal interaction, feedback, nodding, head movements.

1 Introduction

Natural communication does not only include verbal utterances but a wide variety of non-verbal means, ranging from physiological displays to paralinguistic vocalisations and head movement, hand gestures, and body posture. They all have important functions in the communication as whole, as they provide tacit cues of the interlocutors' emotional state and also allow the speaker to control the conversation and manage feedback and turn-taking. Physiological signals can effectively indicate the interlocutor's emotional state, and usually they are unintentional (e.g. we blush for embarrassment, or our pupils dilate for surprise), while hand gesturing, body and head movements, gaze, and facial expressions appear as more controlled means of communication, although they also can indicate the speaker's emotions and focus of attention in a spontaneous and automatic manner. The speakers need not be fully aware about their behaviour, e.g. tilting one's head, beating one's hand, or swaying one's body can be typical behaviours for a speaker, but not necessarily something that the speaker intentionally aims to act like. [1] talks about the degrees of intentionality and awareness in bodily communication, and discusses the concepts of *indicate*, *display*, and *signal* to specify three different degrees of intentional behaviour. "Indicate" denotes the agent's actions lacking conscious intentionality (automatic reactions like blushing), while the two others are associated with greater degrees of awareness and intentionality: "display" refers to the agent showing something intentionally, and

"signal" is a second-order display, implying that the communicator does not only display a meaning, but also their intention that the partner understands their intention to display the meaning. Because of its more spontaneous nature, non-verbal communication is often regarded as more truthful or authentic expression of the speakers' meaning than their verbally expressed utterances. Although it is difficult to identify behavioural cues that would reliably indicate the speaker's truthfulness (or deceit) in general, it is possible to detect non-verbal behaviour patterns that characterize individual speakers as a whole and then to look for changes in their behaviour that can be used as an indication of their emotional, intentional, and attentional state. In fact, such indirect lie detection methods have been successfully used in recent deception studies and they have produced more accurate results than the attempts to identify specific behaviours that are thought to be related with lying. For instance, [8] report that their subjects could distinguish liars from truth-tellers more accurately, if they were asked to identify changes in the people's behaviour rather than explicitly asked to look for liars.

In everyday interactions the processing of multi-modal information is necessary for smooth communication, and much of this socially conditioned. For instance, recognition of social signals affects the interpretation of the partner's message as a humorous or a sarcastic comment, and helps the construction of shared context and mutual understanding. The signals are also important to take into account when planning one's own contribution and intending to coordinate the flow of conversation. Relevant signals in this respect include gesturing to catch the partner's attention to a particular element of interest in the context, turning one's head to the speaker to show one's willingness to be engaged in the interaction, or looking away from the partner, to provide indirect cues of one's non-understanding or lack of interest in the presented message. Such social signalling models are useful for various human-computer applications where the goal is to build more natural interactive systems. We can say that multimodality increases the system's *affordance*, the concept brought to HCI by [21] and suggested by [9] to be used especially with respect to natural language interactive systems: the users need not spend extra time wondering how to operate the interface in order to get their task completed, as multimodal natural language techniques *afford* interaction and lend themselves to the intuitive use of the system.

This paper studies the interlocutors' multimodal activity, such as hand, head and body movement, from the point of view of feedback and construction of shared context. The paper is structured as follows. Section 2 discusses previous studies and sets the scene for multimodal feedback studies. Section 3 presents the corpus of First Encounters, including annotations and our methodology. Section 4 discusses our studies concerning conversational feedback, and reports the results. Section 5 concludes the paper and provides future prospects within intercultural dimensions of the work.

2 Multimodal Feedback

Explicit feedback is important to signal that the speaker's message got through to the partner, but simultaneously it also displays the partner's willingness to maintain good contact and rapport with the speaker. Earlier research has emphasised that multimodal

signals serve social functions by creating bonds and shared understanding but also convey information by reflecting the speakers' attitudes, mood, and emotions [7].

Other important functions deal with their use to regulate the flow of information. For instance, [14] talks about *meta-discursive* function of hand gestures, and shows how different hand forms represent semantic themes which are motivated by different communicative needs on the utterance level, or by communication management. For example, pointing gestures can direct the partner's attention to an important piece of information in the utterance, they can be used to halt the conversation, and they can also mark the next speaker. Although pointing is fairly a distinct gesture, it can also vary in its form: it can be made by an extended finger, an extended hand with open palm, or by a tool such as a pencil that the participant can manipulate in their hand.

Also eye-gaze is an effective means to give and elicit feedback. Gaze indicates where the speaker's focus of attention is directed, and so looking at the conversational partner or looking away from the partner can indicate the partner's understanding of the presented information or willingness to continue interaction. Gaze is also important in indicating if the speaker wishes to keep the turn although hesitant in their wording, or if the speaker wishes to offer the turn to the partner [10]. Mutual gaze is needed to agree on smooth turn-taking and grounding of information [e.g. 4, 16,19].

One of the most important feedback signals is head movement. Nodding is a common way to give acknowledgement and agree with what the speaker says, while side turns effectively signal the change in the participant's focus of attention. [20] compared head nods in three Nordic languages and noticed statistically significant differences in their frequency. [23] discuss nods in Finnish interactions and point out a difference in up-nods and down-nods, the former being mainly used if the speaker presents information that is somehow new to the listener while the latter is neutral acknowledgement of the presentation.

As for the body posture, leaning forward often means interest while leaning backward signals withdrawing from the conversational situation, and they can thus be used to control and coordinate interaction [13]. Some body movements are also used to fill pauses in conversation: the speaker does not want to take the turn or is unable to take the turn. In multiparty interactions, spatial configurations of the participants can show the participants' relation to each other and distinguish the primary and secondary recipients of the speaker [5].

2.1 Cooperation and Shared Context

The interlocutors cooperate with each other and construct a shared context by exchanging information [6, 9]. Communication can thus be seen as cooperative activity in which the interlocutors are engaged in. Cooperation can manifest itself on several levels, from tight task-based collaboration in order to achieve a particular goal, to behaviour patterns that occur simultaneously when the interlocutors interact. An important component of this process is to construct a shared context in which to achieve the shared goal. The construction of the shared context takes place via interactive evaluations of the partner's contributions, whereby the agents give feedback to each other on the current state of communication: if they are willing to continue in the

interaction and what is the level of understanding and agreeing with the partner on the information content exchanged. In the widest sense, feedback refers to the agent's response to the partner's utterance in general, i.e. it is a conscious or unconscious reaction to the changes that take place in the agent's communicative environment. Feedback in this sense is used to refer to the agent's evaluation of the basic enablements for interaction, so it is synonymous to the agent monitoring the interaction in general. Often feedback is understood in a narrower sense as a particular expression used to give feedback to the partner or to elicit feedback from the partner, on some communicatively relevant aspect of interaction. Multimodal feedback has been mainly regarded as displaying certain aspects of the speakers' cognitive and emotional state and thus they allow the interlocutors to monitor each other's emotions and understanding in a natural way.

As mentioned, the interlocutors provide feedback by head, hand, and body, besides explicit verbal feedback. Gestures, facial expressions, and eye-gazing are an unobtrusive means to construct shared context effectively, so that successful interaction can take place. In recent years, a number of studies concerning synchrony and cooperation has increased [12, 17, 18, 22]. Copying of each other's movements, gestures and body postures often occurs in conversations, and this kind of behaviour where the participants align their behaviour with each other can be understood as signalling cooperation between the participants building a common ground. In psycholinguistic and social interaction studies such synchronous behaviour is usually called alignment [22] or mimicry or copying [e.g. 18].

On the other hand, the function of multimodal signals can vary depending on the context. For example, forward leaning can be related to adjusting one's position, but it can also be interpreted as the partner finding the situation uncomfortable and wanting to leave. Backward leaning can display a relaxed participant in a happy listener position, or withdrawal from the conversation. Knowledge of the context is thus a key factor in understanding the function of multimodal signals in interactive situations. [14] argues that the meaning and function of gestures depend on the different relations they have to the surrounding context, i.e. their meaning is created in interaction of the linguistic and dialogue contexts in which they occur, see also [10].

2.2 Referential Schemas

Speakers make frequent verbal and non-verbal references to situation and language context. Their interpretations of linguistic expressions, and of the whole interaction, are influenced by observations from the spatial and visual environment. We can assume that when processing linguistic expressions, human cognition operates on the basis of particular action representations which representations can be defined as general schemas. The schemas can then be employed in different contexts with different multimodal means, and be formulated as scripts or frames or patterns.

Much research is being conducted concerning how senso-motor activity constitutes linguistic representations. Discussion has concerned their origin being based on experience or an innate skill that directs the interpretation. [24]. We emphasise the interaction between linguistic expressions and experience: the meaningful units, either

verbal or non-verbal elements, are created in interaction of the agents with their environment and the other agents. Every action is characterized as goal-oriented, and the signals which have been successfully used in communication to achieve a particular goal are likely to be repeated in future situations too. Their repetitive success reinforces their usefulness. Referential schemas are learned through acting, observation and imitation. They are useful in that they allow the agent to plan their actions and estimate the outcome of their actions: based on their previous experience, the agent can select appropriate actions with the desired outcome, and anticipate the results of their actions.

Concerning feedback, a question is related to the agents' understanding of the relation between language and the physical environment, the embodiment of linguistic knowledge via action and interaction. The relation between high-level communication and the processing of sensory information is under intensive research, but it is obvious that motoric activity accompanies speech, e.g. [14] discusses synchrony of gestures and speech, and how gesture peaks coincide with stress in spoken utterances.

3 Data

The video recordings were collected within the Finnish part of the Nordic project NOMCO [20]. The goal of the project is to study the relation and correlations between speech and multimodal signals in face-to-face communication situations in the Nordic countries (Danish, Finnish, and Swedish data; later on a similar project was also started on Estonian data), and to compare conversational strategies in culturally rather similar yet linguistically different (Finnish vs. Danish and Swedish) contexts. To provide a comparable basis for data analysis, special focus was put on the similar collection setup (non-verbal communication between interlocutors meeting for the first time), and also on the use of a uniform annotation scheme [2], which enables



Fig. 1. Two participants interacting with each other for the first time



Fig. 2. The center camera view of the interactants

similar annotation categories and features to be used across the corpora. The Finnish data consists of a total of 16 interactions between 8 female and 8 male participants, average age about 23 years, each taking part in two separate conversations with different partners. The participants did not know each other in advance, and their task was to talk and to get to know each other. Each encounter was about 6-10 minutes long, and it started when one of the participants entered the video recording room where the other was waiting. Each participant was recorded by a separate video camera, and a third camera was positioned so that it recorded both partners simultaneously. Figure 1 shows screenshots of two participants interacting with each other, and Figure 2 is the corresponding center camera view of them.

The data was then annotated using the Anvil software [15] and the NOMCO annotation scheme [2], with respect to head, hand, and body movements that were considered carrying communicative functions. Of all the 645 head movements, 576 (89 %) were related to feedback: 375 gave feedback and 201 elicited feedback. Of all the 402 hand movements, 295 (73%) occurred simultaneously with feedback: 90 to give feedback and 205 to elicit feedback. Finally, of the 96 body movements, 68 (71%) were feedback related: 38 gave feedback and 30 elicited feedback. The other movements were related to turn management or other unspecified functions. Annotation features and their statistics are given in Table 1.

Table 1. Frequency count of head, hand, and body annotation features

Head movements	Count	Hand movements	Count
Backward	38	Handedness	
Forward	77	Both hands	265
Nod (up and down)	345	One hand	137
Tilt	95	Hand movement repetition	
TurnSide	76	Repeated	174
Waggle	11	Single	220
Other	3	Other	8
Total (head)	645	Hand movement interpretation	
Body movement		Deixis	8
Backwards	15	Emphasis	19
Forward	37	Rhythm	185
LeaningAwayFromPartner	41	Standup	2
Other	3	Not classified	188
Total (body)	96	Total (hand)	402

Annotation was done by two annotators independently and checked by an expert annotator. Inter-coder agreement between the two annotators on annotation categories is shown in Table 2. Kappa (κ) is calculated using Anvil's coding agreement facility which automatically compares two annotation files frame by frame (frame = 0.4s) and calculates inter-coder agreement with respect to the joint segmentation and categories, and the overall agreement. The annotators showed very good agreement on body annotation, good or fair agreement on face features, and almost perfect agreement on hand category agreement. The surprisingly low agreement on hand segmentation is obviously due to the annotators' difference in deciding on the start and end points of complex hand gestures, and whether these are classified as one or many gestures.

Table 2. Kappa and %-agreement of some of the annotation categories

Track	Segmentation		Category		Overall	
	κ	%	κ	%	κ	%
Face	0,48	74	0,71	79	0,65	72
Hand	0,09	41	0,95	98	0,11	41
Body	0,86	93	0,79	84	0,91	93

4 Results

Simple correlations were calculated on the basis of the data, to find out how hand, head, and body movement correlate with the feedback function. Detailed analyses are discussed in [13] while here we focus on the combined effects of the different modalities on feedback. These are shown in Figure 3. The first four columns on the left concern body movement, the six last ones on the right concern head movement, and the 10 columns in the middle describe hand movements. The columns are normalised with respect to the frequency counts in Table 2.

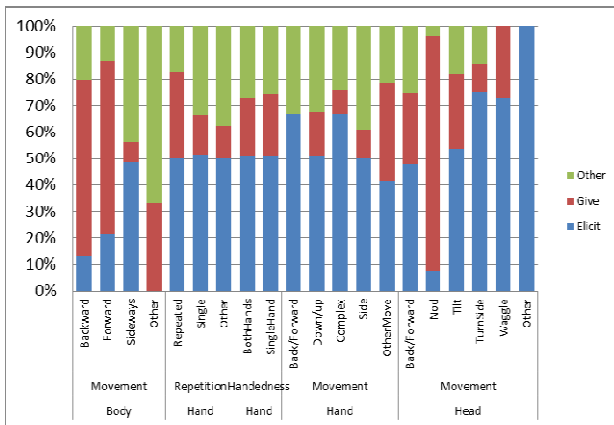


Fig. 3. Give and Elicit feedback in different modalities. Other refers to movements not related to feedback.

We can see that back/forward body movements are mainly used to give feedback and sideways movements to elicit feedback. Sidway movements rarely function to give feedback, but they can also occur in other functions not related to feedback at all. However, other, unspecified body movement types can also function to give feedback. Hand is mainly used to elicit feedback, i.e. it is used to engage the partner into the conversation. In particular, back/forward gestures are related to feedback elicitation. Concerning repetition of hand gestures, almost 40 % of the repeated gestures but only about 20% of single hand gestures give feedback, i.e. twice as many repeated gestures function as feedback giving signals as single gestures. As for handedness, independently from whether the gestures are produced by single hand or both hands, a similar pattern appears: feedback elicitation is twice as common as feedback

giving. Also, one third of all the gestures are used for other functions than providing feedback.

Head movements are also used to provide feedback, and in general they show a similar relative distribution between feedback giving and eliciting functions as hand movements: about third of the movements function as feedback giving. Back/forward movement, tilting, side turns, and waggle are more common in eliciting feedback, but nodding, however, is almost exclusively used to give feedback. Nodding can also be divided into up-nods and down-nods depending on which way the movement starts. Although this may not always be clear, in [23] it was noticed that the difference is related to the interpretation of nodding: up-nods are used when the speaker has presented information that is new to the listener in the given context, while down-nods signal neutral acknowledgement of the information that is expected in the context.

Most commonly feedback is related to spoken feedback particles and backchanneling vocalisations, and it is interesting to see correlations between verbal and non-verbal communication. The four most common verbal feedback signals in the Finnish corpus are *joo* (yeah), *nii(n)* (yes but), *okei* (okay), *aivan* (exactly). The difference between *joo* and *nii* is related to the novelty value of the presented information: *joo* indicates that the speaker acknowledges the presented information, while *nii(n)* indicates that the speaker has some reservations about it, or can add more to it. *Okei* and *aivan* indicate the speaker's stronger agreement or commitment to the presented information. Figure 3 depicts co-occurrence frequency counts of the common verbal feedback and head movements, normalised with respect to time (in our case: one minute). We notice that nodding is common with both the neutral acknowledgement (*joo*) and the acknowledgement with reservations (*nii*), and in fact, the difference appears to be related to the direction of nodding as discussed in [23]: the neutral *joo* most often co-occurs with down-nods, while *nii* usually co-occurs with up-nods.

Back- and forward head movement is about twice as common with *nii* than with *joo* or *okei*, and also tilting of the head co-occurs more than twice as often with *nii* as with *joo* or *aivan*. We can thus assume that tilting is related to a surprise or novelty value of the presented information, and that the listener has some reservations about it that prevent the uptake of the information as such. Side turns, however, generally signal change of attention away from the partner, and this may explain why they are not common with any of the four feedback particles. They are extremely rare with *aivan* (exactly), which is understandable as the speaker would verbally express strong agreement with the presented information but non-verbally display contradiction or non-interest with it. Verbal and non-verbal feedback activity tend to be produced, and also get interpreted, as semantically aligned in smooth communication

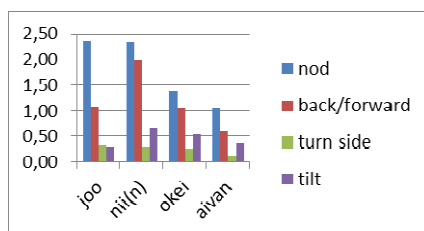


Fig. 4. Frequent co-occurrences of verbal feedback and head movement signals

5 Conclusion and Future Prospects

This paper has focussed on the head, hand, and body movements and their use in conversational interactions as means of visual interaction management. It is clear that hand gestures, head movement and body posture are important multimodal means for interaction management and for providing feedback of the current state of the interaction. We presented some observations of their co-occurrence and correlations in feedback giving and eliciting situations, and also provided interesting co-occurrence data of verbal and non-verbal feedback expressions.

These observations will be used for further multimodal and multicultural studies. We will investigate visual interaction management as part of human-computer interaction, and focus on the role of multimodal signals in the control and coordination of interaction. Experiments will concern the participant's engagement in conversational interactions, and we will use various features, especially multimodal, besides verbal features, to measure their conversational activity. We will also build models of the multimodal strategies that the interlocutors have at their disposal to construct shared understanding and to advance their goals, to be applied to interactive applications.

Since the NOMCO first encounter corpus is analogous to the other similar corpora collected at the other Nordic countries, it is possible to compare interaction strategies in different (cultural) contexts, especially in cultures and among languages which are closely related. Such comparison will allow us to deepen our understanding of the various multimodal signals and their impact and function in interactions as well as in intercultural communication.

References

1. Allwood, J.: An Activity Based Approach to Pragmatics. In: Gothenburg Papers In Theoretical Linguistics 76. Department of Linguistics. Göteborg University (2000)
2. Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., Paggio, P.: The MUMIN Coding Scheme for the Annotation of Feedback, Turn Management and Sequencing Phenomena. Multimodal Corpora for Modelling Human Multimodal Behaviour. Special Issue of the International Journal of Language Resources and Evaluation 41(3-4), 273–287 (2007)
3. André, E., Pelachaud, C.: Interacting with Embodied Conversational Agents. In: Jokinen, K., Cheng, F. (eds.) *Speech-based Interactive Systems: Theory and Applications*. Springer (2009)
4. Argyle, M., Cook, M.: *Gaze and Mutual Gaze*. Cambridge University Press (1976)
5. Battersby, S.: *Moving Together: the organization of Non-verbal cues during multiparty conversation*. PhD Thesis, Queen Mary, University of London (2011)
6. Clark, H.H., Schaefer, E.F.: Contributing to Discourse. *Cognitive Science* 13, 259–294 (1989)
7. Feldman, R.S., Rim, B.: *Fundamentals of Nonverbal Behavior*. Cambridge University Press (1991)
8. Hart, C.L., Fillmore, D.G., Griffith, J.D.: Indirect Detection of Deception: Looking for Change. *Current Research in Social Psychology* 1(9), 134–142 (2009)

9. Jokinen, K.: Rational communication and affordable natural language interaction for ambient environments. In: Lee, G.G., Mariani, J., Minker, W., Nakamura, S. (eds.) *IWSDS 2010*. LNCS, vol. 6392, pp. 163–168. Springer, Heidelberg (2010)
10. Jokinen, K.: Pointing gestures and synchronous communication management. In: Esposito, A., Campbell, N., Vogel, C., Hussain, A., Nijholt, A. (eds.) *COST 2102 Int. Training School 2009*. LNCS, vol. 5967, pp. 33–49. Springer, Heidelberg (2010)
11. Jokinen, K., Furukawa, H., Nishida, M., Yamamoto, S.: Gaze and Turn-taking behaviour in Casual Conversational Interactions. *ACM Trans. Interactive Intelligent Systems* (2013)
12. Jokinen, K., Pärkson, S.: Synchrony and copying in conversational interactions. In: *Proceedings of the 3rd Nordic Symposium on Multimodal Communication*, vol. 15, pp. 18–24 (2011)
13. Jokinen, K., Wilcock, G.: Multimodal Signals and Holistic Interaction Structuring. In: *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, Mumbai, India (2012)
14. Kendon, A.: *Gesture. Visible Action as Utterance*. Cambridge University Press (2005)
15. Kipp, M.: Anvil - A Generic Annotation Tool for Multimodal Dialogue. In: *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)*, pp. 1367–1370 (2001)
16. Lee, J., Marsella, S.C., Traum, D.R., Gratch, J., Lance, B.: The Rickel Gaze Model: A Window on the Mind of a Virtual Human. In: Pelachaud, C., Martin, J.-C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) *IVA 2007*. LNCS (LNAI), vol. 4722, pp. 296–303. Springer, Heidelberg (2007)
17. Levitan, R., Gravano, A., Hirschberg, J.: Entrainment in Speech Preceding Back-channels. In: *Proceedings of ACL 2011*, pp. 113–117 (2011)
18. Mancini, M., Castellano, G., Bevacqua, E., Peters, C.: Copying Behaviour of Expressive Motion. In: Gagalowicz, A., Philips, W. (eds.) *MIRAGE 2007*. LNCS, vol. 4418, pp. 180–191. Springer, Heidelberg (2007)
19. Nakano, Y., Nishida, T.: Attentional Behaviours as Nonverbal Communicative Signals in Situated Interactions with Conversational Agents. In: Nishida, T. (ed.) *Engineering Approaches to Conversational Informatics*. John Wiley (2007)
20. Navarretta, C., Ahlsén, E., Allwood, J., Jokinen, K., Paggio, P.: Feedback in Nordic First-Encounters: a Comparative Study. In: *Proceedings of the Language Resources and Evaluation Conference (LREC-2012)*, Istanbul, Turkey (2012)
21. Norman, D.A.: *The Psychology of Everyday Things*. Basic Books, New York (1988)
22. Pickering, M., Garrod, S.: Towards a mechanistic psychology of dialogue. *Behavioral and Brain Sciences* 27, 169–226 (2004)
23. Toivio, E., Jokinen, K.: Multimodal Feedback Signaling in Finnish. In: *Proceedings of the Human Language Technologies – The Baltic Perspective (2012)*; Published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License
24. Tomasello, M.: *First verbs: A case study of early grammatical development*. Cambridge University Press, Cambridge (1992)

Keyboard Clawing: Input Method by Clawing Key Tops

Toshifumi Kurosawa, Buntarou Shizuki, and Jiro Tanaka

University of Tsukuba, Japan
{kurosawa, shizuki, jiro}@iplab.cs.tsukuba.ac.jp

Abstract. We present a directional and quantitative input method by clawing key tops, Keyboard Clawing. The method allows a user to input a direction and quantity at the same time without moving his/her hands much from the keyboard's home position. As a result, the user can seamlessly continue typing before and after inputting the direction and quantity. We found that clawing direction is classified using clawing sounds with an accuracy of 68.2% and that our method can be used to input rough quantity.

Keywords: keyboard, acoustic sensing, gesture, input method.

1 Introduction

Some researchers have proposed input methods using keyboards in order to provide seamless inputs while typing. Block et al. presented a touch-display keyboard [1], which recognizes user's hands movements by touch sensors attached under each key top and provides a graphical display on its surface. This allows a user to use the keyboard like a touch-display. Tsukada et al. suggested a PointingKeyboard [2], which allows users to perform pointing on the keyboard by recognizing the user's hands position using an IR sensor attached on the keyboard. However, these approaches need keyboards to be modified significantly, thus they are expensive to realize.

Unlike these approaches, Keyboard Clawing, our input method, uses *sounds caused by clawing key tops* to detect a user's operation to allow four-directional input with quantity. It recognizes the direction in which the user claws key tops and distance of clawing by analyzing the sounds. This method has the following three advantages.

1. A user can input a direction and quantity at the same time using a keyboard.
2. A user can begin the input fast because the user can claw key tops without moving his/her hands much from the home position.
3. The system is inexpensive to realize because it only needs one microphone to be attached to a keyboard the a user is accustomed to using.

We begin with the previous work on input methods by extending keyboards' input capability or using acoustic sensing. Next, we explain the design and implementation of our Keyboard Clawing and its applications. Finally, we refer to a user study that shows users' performance of our method.

2 Related Work

There are some input methods using keyboards other than mentioned above. Dietz et al. proposed a practical pressure-sensitive keyboard [3] that could detect pressure information of key pressing. Fallot-Burghardt et al. presented an extended keyboard, Touch&Type [4], which allows users to use the keyboard surface for touchpad-like pointing. Rekimoto proposed the ThumbSense [5]. ThumbSense senses contact of user's finger to the touchpad. Under this condition, a user can click by pressing a normal key (e.g., *F*) instead of click buttons. This allows a user to click without moving his/her hands from the keyboard's home position. Keyboard Clawing also extends keyboards' input capability. In contrast to these previous approaches, our method uses clawing sounds on the keyboard to detect users' operations.

Many approaches using acoustic sensing to detect users' operations have been proposed [6–10]. Keyboard Clawing also uses acoustic sensing. Our method differs from these approaches in terms of using *clawing sounds on the keyboard* to detect a user's operation.

Researchers have also classified the sounds to detect gestures on keyboards. Zhuang et al. [11] and Berger et al. [12] distinguish pressed keys by analyzing typing sounds. Kato et al. presented Surfboard [13], which allows a user to perform two-directional input without quantity. Surfboard uses sounds produced by a user's large hand movements on the keyboard. Since Surfboard only needs a microphone, it is realized at low cost. Keyboard Clawing also uses sounds on keyboards to detect gestures on them but provides four-directional and quantitative input to users.

3 Keyboard Clawing

In this section we explain how to use Keyboard Clawing. A user puts his/her hands in a keyboard's home position and claws the key tops with his/her finger to up, down, right, or left (Figure 1). If a user wants to input up or down, he/she claws the key tops in either direction with his/her index finger. If a user wants to input right or left, he/she claws the key tops in either direction with his/her thumb. Note that this design intends to reduce the movement of the user's hand from a keyboard's home position.

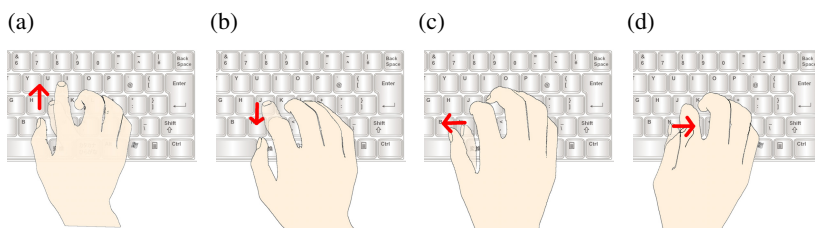


Fig. 1. How to claw key tops in Keyboard Clawing. If a user wants to input up or down, he/she claws key tops to each direction by his/her index finger (a, b). If a user wants to input right or left, he/she claws key tops to each direction by his/her thumb (c, d).

The system uses the clawing sounds for detection of clawing direction and distance. Specifically, the distance is the number of grooves between keys the user claws. A user measures the clawing distance at a rough estimate to input quickly. For example, when the user claws *J* to *N*, the system recognizes the input as “1-down”. When the user claws *M* to *B*, the system recognizes the input as “2-left”.

4 Implementation

To recognize how a user claws key tops using the clawing sounds, we attach a piezo microphone (Shadow SH-710, Figure 2) on a keyboard (Dell KB212-B, Japanese keyboard) to capture the sounds (sampling rate: 96 kHz, quantization bit rate: 16 bit) as shown in Figure 3. The system first scales the captured sounds from -1.0 to 1.0 and then reduces noise of the scaled sounds using MMSE-STSA [14]. When the system detects a peak exceeding a threshold (0.3), it regards the sound as a clawing one and recognizes the clawing direction and distance by analyzing the sounds. The system also monitors the key release events in order to distinguish the clawing sounds from typing sounds. If the system detects the sounds 500 ms after the key release events occur, it regards the sounds as typing sounds, not clawing ones.



Fig. 2. Piezo microphone

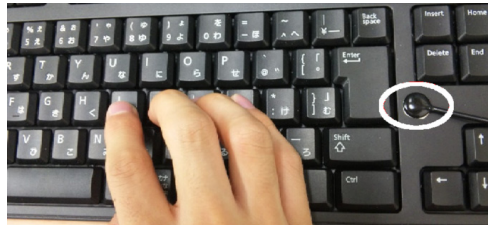


Fig. 3. Piezo microphone on a keyboard

4.1 Detecting the Direction

We use frequency distribution of the clawing sounds in order to detect the clawing direction. The system obtains a frequency distribution of the 2000 samples (20.8 ms) centered of the sounds’ first peak using the Fast Fourier Transform (FFT). The FFT frame size is 2000. This results in 1000 feature values (band: 0 - 48 kHz). The clawing sounds are classified using the feature value and a support vector machine (SVM) as a discriminator.

4.2 Detecting the Distance

We use the number of peaks of the clawing sounds in order to detect the clawing distance because a peak occurs when the user claws key tops and the finger passes on a groove between keys. We first extract 48000 samples (500 ms) and transform the sounds by taking the absolute amplitude of each sample. Then, we compute the simple moving average of each transformed sample for 50 periods (Figure 4d) to remove stray peaks

and smooth the waveform. Next, the system scans the smoothed waveform and counts the number of samples that exceed threshold (0.12) (Figure 4e). In scanning, the system ignores peaks within 2000 samples (20.8 ms) from a sample exceeding the threshold because the peaks are considered to be caused by clawing the same groove of the previous peak.

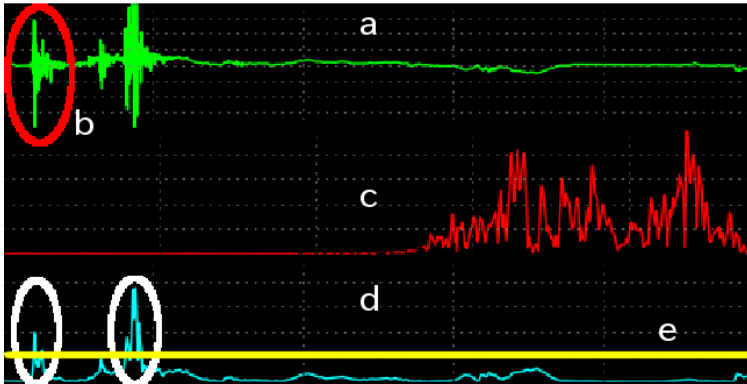


Fig. 4. (a) Example clawing sounds waveform. (b) First peak used for direction detection. (c) Frequency distribution of the example sounds. (d) Smoothed waveform. (e) Threshold of distance.

5 Applications

We used Keyboard Clawing for shortcuts in word processing, in which direction and quantity are frequently input.

Text selection and changing the font size: We use our Keyboard Clawing to select text and change the font size. Table 1 shows an example of functions assigned to our method. Figure 5 shows how one uses our method.

Scrolling: We also apply our Keyboard Clawing to scrolling. Table 2 shows an example of functions assigned to our method. Figure 6 shows how one uses our method. A user cannot only adjust input quantity but also change function by clawing distance (e.g., 1-down: scroll a line. 2-down: scroll a page).

In these cases, a user can select text or change the font size or scroll without moving his/her hands from the keyboard’s home position, unlike keyboard shortcuts or pointing devices, so that he/she can seamlessly continue typing before and after the operation.

Table 1. Example of functions assigned to Keyboard Clawing

Direction\Distance	1	2	3
Up	font size +5pt	font size +10pt	font size +15pt
Down	font size -5pt	font size -10pt	font size -15pt
Left	select a left character	select a left word	select to top of the line
Right	select a right character	select a right word	select to end of the line

Table 2. Example of scroll functions assigned to Keyboard Clawing

Direction\Distance	1	2	3
Up	1 line up	1 page up	to top
Down	1 line down	1 page down	to end

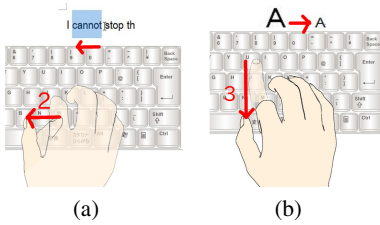


Fig. 5. Text selection and changing font size by Keyboard Clawing. (a) Select a left word. (b) Scale-down 15pt.

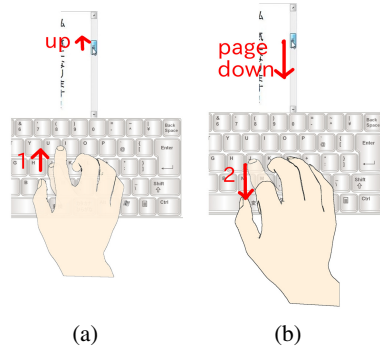


Fig. 6. Scrolling by Keyboard Clawing. (a) Upwards 1 line. (b) Downwards 1 page.

6 Evaluation

To examine the accuracy of detecting clawing direction and distance, we conducted a user study in a quiet room. We recruited 12 volunteer participants (11 male and one female) aged from 22 to 24 years. One participant made too many mistakes in the study, thus we use data from the other 11 participants. Their nails were of ordinary length.

Before the study, participants were told how to use Keyboard Clawing and practiced till they were accustomed to the method. Participants were able to ask an experimenter about the method or the study during the practice. To prevent loss of concentration, they could also take a rest between tasks freely.

In a trial, the participants were told both direction and quantity. The quantities were 1-4 in clawing to up or down, and 1-10 in clawing to left or right. Each participant conducted trials five times in each direction and quantity and thus performed $(4 \times 2 + 10 \times 2) \times 5 \times 11 = 1,540$ trials. We only showed visual feedback to inform participants that the system had detected the sounds. Note that Keyboard Clawing is intend to allow users to input quickly, thus we suppose that when one uses the method he/she measures the clawing distance as a rough estimate. Hence, we told participants to follow the indicated quantity roughly.

7 Results and Discussion

7.1 Accuracy of the Distance

Table 3 shows the accuracy of the distance. The results show low accuracy (between 19.6% - 39.5%). There are two reasons for this. First, when the clawing finger touches a key top, a peak occurs if the user touches it strongly but not if the user touches it softly. Second, participants follow the indicated distance roughly.

To study the accuracy further, we adopt a tolerance. In clawing to up and down, the accuracies were 98.2% and 90.5%, respectively, when the tolerance was two. Similarly in clawing to left and right, the accuracies were 91.5% and 89.6%, respectively, when the tolerance was three. Thus, the results are highly accurate for the tolerances of two and three. Hence our method is applicable to inputting a rough quantity.

Table 3. Accuracy of the distance (%)

Direction \ Tolerance	Tolerance				
	0	1	2	3	4
Up	39.5	83.6	98.2	100.0	100.0
Down	37.3	75.0	90.5	100.0	100.0
Left	24.4	58.5	80.2	91.5	96.9
Right	19.6	54.7	76.2	89.6	95.6

7.2 Accuracy of the Direction

To obtain the accuracy of the direction, we extracted the feature of the participants' clawing sounds and then conducted a 35-fold cross-validation. To classify, we use a SVM provided by the Weka Machine Learning toolkit [15]. The SVM type was C-SVC, the kernel type was linear kernel, the cost value was 512.0, the gamma value was 0.000125, and we set "Normalize" true.

Table 4. Accuracy of direction of the per-user test (%)

Input \ Classified as	Classified as				Accuracy
	Up	Down	Left	Right	
Up	60.5	14.1	15.9	9.5	60.5
Down	12.7	62.3	13.2	11.8	62.3
Left	4.9	2.7	72.6	19.8	72.6
Right	2.4	3.5	16.7	77.4	77.4
Average	-	-	-	-	68.2

Table 5. Accuracy of direction of the cross-user test (%)

Input \ Classified as	Classified as				Accuracy
	Up	Down	Left	Right	
Up	41.8	22.7	23.2	12.3	41.8
Down	24.5	46.4	12.3	16.8	46.4
Left	10.0	7.5	56.9	25.6	56.9
Right	4.4	4.5	25.3	65.8	65.8
Average	-	-	-	-	56.4

In per-user test, we conducted the cross-validation to each participant's feature value. Table 4 shows detailed results of the per-user test. The average in all directions is 68.2%. Table 5 shows a confusion matrix of cross-user test. In the cross-user test, we conducted the cross-validation to all participants' feature values. The average in all directions is 56.4%. This result shows we can classify the clawing directions by frequency distribution of the sounds with accuracy of about 70%. The result of cross-user test was 56.4%.

11.8 points lower than that of the per-user test. This means that the clawing sounds differ among users. Hence, before using our method actually, users need calibration. We also found that the accuracy of left or right clawing is higher than that of up or down clawing. This is because left or right clawing is louder than up or down clawing, thus the features of sounds appear clearly.

7.3 The Number of Key Pressing in Clawing

To study whether it is rational to monitor key release events in order to distinguish the clawing sounds from typing sounds, we also examine *the number of key pressings* (*NKP*) in clawing. Table 6 shows detailed *NKP*, and Figure 7 shows maps of pressed keys in each directional clawing in the study. In these maps, the more times a key is pressed in clawing, the darker the key is colored. The results show participants tended to press keys when clawing large distances and in the latter part of clawing, thus the system failed to distinguish clawing sounds from typing sounds. To solve this problem, we will use the previous sounds. Even if key release events occur, the system will regard the sound as a clawing one if the clawing sounds occurred just previously.

Table 6. Detailed *NKP* (%)

Direction \ Distance	Distance										Average
	1	2	3	4	5	6	7	8	9	10	
Up	1.8	0.0	3.6	1.8	-	-	-	-	-	-	1.8
Down	0.0	0.0	0.0	0.0	-	-	-	-	-	-	0.0
Left	0.0	0.0	1.8	3.6	5.5	7.3	14.6	9.1	16.4	18.2	7.6
Right	3.6	1.8	1.8	0.0	1.8	3.6	1.8	9.1	10.9	7.3	4.2
Average	1.4	0.5	1.8	1.4	3.6	5.5	8.2	9.1	13.6	12.7	4.5

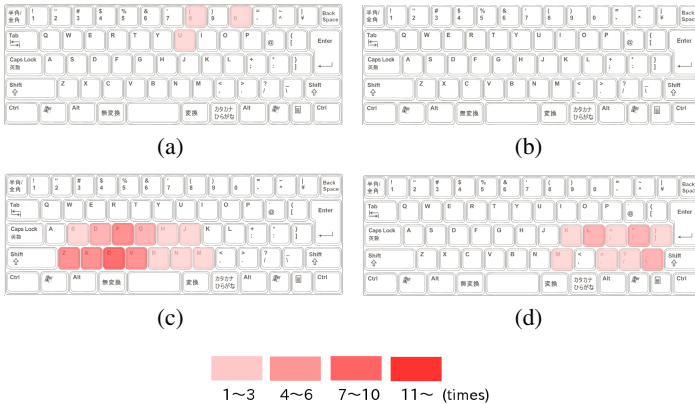


Fig. 7. Maps of pressed keys in each directional clawing. (a) Up. (b) Down. (c) Left. (d) Right.

8 Conclusions and Future Work

We presented a directional and quantitative input method by clawing key tops, Keyboard Clawing. The system detects the sounds caused by clawing key tops without user's hands moving much from the keyboard's home position. Keyboard Clawing enables a user to continue typing seamlessly before and after clawing. We conducted a user study and found the accuracy of the direction to be 68.2% for each participant, thus the algorithm needed to be improved. Additionally, the accuracy of distance is approximately 90% with a tolerance of two or three, thus our method is applicable to inputting a rough quantity.

For future work, we want to improve the detecting direction algorithm. We have two plans for new algorithms. In one, although we used only the first peak to detect the direction, we will apply the SVM classifier to all peaks in a clawing sound and then adopt a direction that detected the most as an input direction. In the other, we will adopt time series analysis using Hidden Markov Models. After improving the algorithm, we also want to compare our method with other input methods (e.g., keyboard shortcuts, mouse pointing) in terms of input speed and error rate to confirm the effectiveness of Keyboard Clawing.

References

1. Block, F., Gellersen, H., Villar, N.: Touch-display keyboards: transforming keyboards into interactive surfaces. In: Proceedings of the 28th International Conference on Human Factors in Computing Systems, CHI 2010, pp. 1145–1154 (2010)
2. Tsukada, Y., Hoshino, T.: PointingKeyboard: An input device for typing and pointing. In: Proceedings of Workshop on Interactive Systems and Software, WISS 2002, pp. 999–99 (2002) (in Japanese)
3. Dietz, P., Eidelson, B., Westhues, J., Bathiche, S.: A practical pressure sensitive computer keyboard. In: Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology, UIST 2009, pp. 55–58 (2009)
4. Fallot-Burghardt, W., Fjeld, M., Speirs, C., Ziegenspeck, S., Krueger, H., Läubli, T.: Touch&type: a novel pointing device for notebook computers. In: Proceedings of the 4th Nordic Conference on Human-Computer Interaction: Changing Roles, NordiCHI 2006, pp. 465–468 (2006)
5. Rekimoto, J.: ThumbSense: automatic input mode sensing for touchpad-based interactions. In: CHI 2003 Extended Abstracts on Human Factors in Computing Systems, CHI EA 2003, pp. 852–853 (2003)
6. Murray-Smith, R., Williamson, J., Hughes, S., Quaade, T.: Stane: synthesized surfaces for tactile input. In: Proceedings of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems, CHI 2008, pp. 1299–1302. ACM (2008)
7. Harrison, C., Hudson, S.: Scratch input: creating large, inexpensive, unpowered and mobile finger input surfaces. In: Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology, UIST 2008, pp. 205–208 (2008)
8. Harrison, C., Schwarz, J., Hudson, S.: TapSense: enhancing finger interaction on touch surfaces. In: Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, UIST 2011, pp. 627–636 (2011)
9. Lopes, P., Jota, R., Jorge, J.: Augmenting touch interaction through acoustic sensing. In: Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces, ITS 2011, pp. 53–56 (2011)

10. Harrison, C., Xiao, R., Hudson, S.: Acoustic barcodes: passive, durable and inexpensive notched identification tags. In: Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology, UIST 2012, pp. 563–568 (2012)
11. Zhuang, L., Zhou, F., Doug, T.: Keyboard acoustic emanations revisited. *ACM Transactions on Information and System Security* 13(1), 1–3 (2009)
12. Berger, Y., Wool, A., Yeredor, A.: Dictionary attacks using keyboard acoustic emanations. In: Proceedings of the 13th ACM Conference on Computer and Communications Security, CCS 2006, pp. 245–254 (2006)
13. Kato, J., Sakamoto, D., Igarashi, T.: Surfboard: keyboard with microphone as a low-cost interactive surface. In: Adjunct Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology, UIST 2010, pp. 387–388 (2010)
14. Ephraim, Y., Malah, D.: Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Transactions Acoustic Speech & Signal Processing* 32, 1109–1121 (1984)
15. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. *Explorations Newsletter* 11(1) (2009)

Finger Controller: Natural User Interaction Using Finger Gestures

Unseok Lee and Jiro Tanaka

Department of Computer Science, University of Tsukuba,
Tennodai, Tsukuba, 305-8577 Ibaraki, Japan
{leeunseok, jiro}@iplab.cs.tsukuba.ac.jp

Abstract. We present a new natural user interaction technique using finger gesture recognition and finger identification with Kinect depth data. We developed a gesture version drawing, multi-touch and mapping on 3d space interactions. We implemented three type interfaces using their interaction such as air-drawing, image manipulation and video manipulation. In this paper, we explain finger gesture recognition method, finger identification method and natural user interactions in detail. We show the preliminary experiment for evaluating accuracy of finger identification and finger gesture recognition accuracy, evaluating user questionnaire for interaction satisfaction. Finally, we discuss the result of evaluation and our contributions

Keywords: NUI, Human Computer Interaction, Finger Gesture Recognition, Finger Identification.

1 Introduction

The interaction using hand gestures is a popular field in Human Computer Interaction(HCI) and consequently many related research papers have been proposed. Some of them propose media player manipulation interaction using hand motion gesture [3] and glove-based hand gesture interaction. Vision-based hand gesture recognition system was proposed as well [5], [8]. However, the research approaches that propose the mounting of additional device on the body [6] are usually troublesome and not natural. The vision-based hand gesture recognition is not practical for robust hand gesture recognition because of much influence by light and background clutter. Recently, new horizons are open to the HCI field with the development of sensors and technology [2] such as Kinect, DepthSense and Leap motion. This development has made possible robust recognition, like finger gesture recognition in bad conditions such as dark light and rough background. This depth-based sensor and technology provide a robust recognition, but many research works using them do not provide proper natural interaction [1], [4]. They mostly provide simple hand gesture interactions or hand motion interactions. They do not provide a natural interaction. On the other hand, Finger gesture recognition with finger identification can provide more practical experience and natural interaction than hand gesture and hand motion [2].

In this paper, we propose a new natural user interaction technique using finger gestures with finger identification. It is called Finger Controller. We implemented three types of interface such as air-drawing, image manipulation and video manipulation using designed natural user interactions. Figure1 shows the interfaces overview.

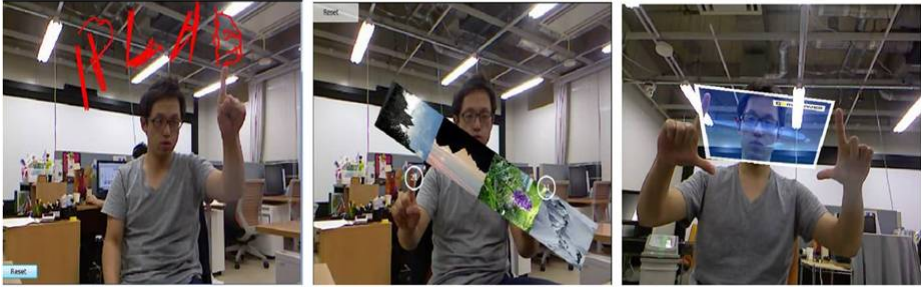


Fig. 1. Air-Drawing(Left), Image Manipulation(Middle), Video Manipulation(Right)

2 Related Work

Many hand gesture recognition techniques are proposed to interact with objects for natural user interface(NUI). They are implemented using various devices and techniques. Vision based, glove based and depth based are widely used in hand gesture recognition [2]. Vision based hand gesture recognition techniques and interfaces are continuously being proposed.

Chu et al [5] presents self-portrait interface using vision based hand motion gesture. They provide interesting interaction to control camera by the user. However, the recognition is difficult in bad conditions, such as darkness, because they use skin color segmentation. In addition, system functions are limited when user moves far from the camera because they do not use depth data for recognition. Kenn [6] presents an interface for wearable computing applications using glove based finger recognition. His paper implemented finger identification for hand gesture. The research provides rich user experience based on accurate recognition rate. However, it is not practical in all situations because the user cannot always carry a glove. The glove is both unaesthetical and heavy. This condition is not suitable for natural interaction.

Yang et al, [3] proposed a hand motion gesture recognition system using Kinect depth data. They designed hand motion gestures like wave, forward/backward, move up/down, left/right in a media player application. The designed gesture interaction session is performed based on assumptions of the user's intentions. The system used a 3D feature vector for examining the hand trajectory and HMM for hand gesture recognition. This system shows the possibility of using Kinect for hand motion gesture recognition in a contact-less UI. However, the implemented system was not able to recognize fingertip. Therefore, they were not able to provide delicate and natural gestures, such as pinch gesture, spread gesture and flick finger. Raheja et al, [4] proposed a method to recognize and track fingertips and center of palm using Kinect. They track the fingertip by calculating depth image segmentation of hand regions.

Then the palm of the hand would be subtracted from the depth image, so that only the fingers are left. Under most situations, fingertip is minimum depth in each finger. The proposed method provides robust fingertip recognition rate. However, they don't implement any interactions, gesture and finger identification. Thus, this approach is not suitable for NUI. In short, most of the proposed hand gesture interaction techniques have limitation for natural interaction using finger gesture.

3 System Overview

In this section, we will discuss the system hardware we used. Fingertip tracking and finger identification method to detect interactions using finger gesture will be discussed as well.

3.1 Hardware

Our system is comprised of Microsoft Kinect for Xbox 360 sensor and large display. The kinect sensor is used with input data for capturing user gestures using CMOS camera(640x480 pixels) with 30 FPS and Depth camera. For improving system performance, we limit the distance from hands to sensor. We found the appropriate value of distance(0.5m to 0.8m) experimentally.

The large display that we used has a size of 30 inches. The display shows interaction feedback from the implemented application using user gesture.

3.2 Depth Based Fingertip Tracking Method

Our system uses depth data information for the robust finger tracking of finger gesture for the natural interaction. The algorithm designed for the finger tracking is described in the following.

The first step is to gather tracked hands depth data from the kinect sensor, and separate hands from the depth image background. A simple k-means clustering algorithm is implemented for dividing hand point of a frame into cluster. Second, the system makes the hand's contour using a modified Moor Neighbor Tracing algorithm, and compute convex hull by Graham Scan algorithm on detected depth cluster data. Then, the system combines contour information and point in convex hull. Third, for each point in the hull that is a candidate for fingertips, we find the nearest point in the contour curve. We call this set P. For each point in P, we take the two supporting points(P1, P2) in two opposite directions along the contour. P1 and P2 positions are detected by analyzing the hand contour shape pattern, because most hand shapes are similar. Then, the system determines whether these three point(point in P, P1 and P2) is aligned or not. Because we can ensure that it is not fingertip in the case of aligned, system doesn't need to calculate. If these three points are not aligned, the system calculates the distance from midpoint of P1 and P2 to the point in P. The distance is a certain value that was found experimentally, the candidate point in P is a fingertip point (see figure 2(c)). Figure 2-(d) shows that the distance is smaller than a certain value, thus that it is not a fingertip point.

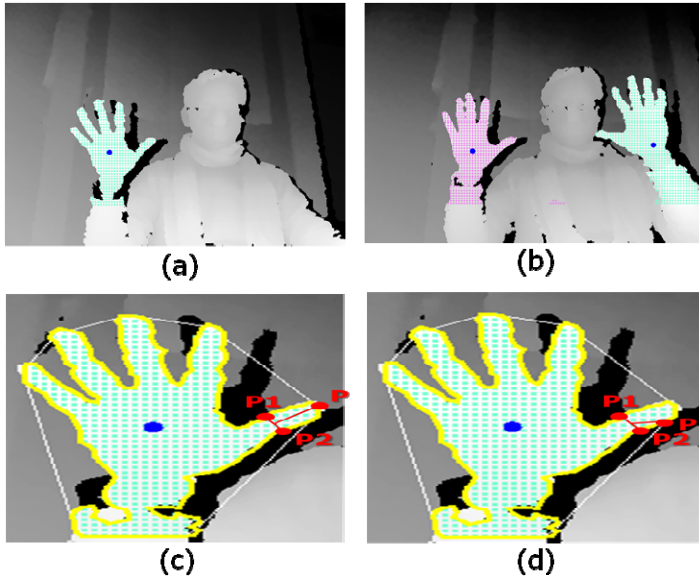


Fig. 2. Hand and Fingertip Detection. (a) Hand Detection using clustering (b) Two hands Detection using clustering (c) Fingertip (d) Not Fingertip.

3.3 Depth Based Finger Identification Method

In this section, we describe a finger identification method for finger gesture recognition. Our system makes interaction based on pre-defined finger gestures such as pinch, spread and rotate gesture, so that the finger identification is important for such delicate recognition. We have motivation of developing identification method that when the people manipulate object such as smart pad and phone devices, they are mainly used thumb and index finger for interaction. Our system tracks fingertip using the method explained in 3.2, then calculates mathematically their coordinates and depth value for finger identification. The method is mainly divided into three parts, described as follows.

The first step is counting the number of fingers. According to this number, the system determines whether it will start to calculate or not. The algorithm starts to calculate when all fingers are extended. Second, we identify the thumb and the index finger. The system uses the distance from the device and shape bases matching for identifying the thumb. In general, among all fingers the thumb has the shortest distance from the device when all fingers are extended. We use hand shape matching for more accurate thumb identification. The thumb can be found more accurately by combining these results, and then we make pairs of all neighbor fingers. System calculates largest distance among one to four (see figure 3(a)). The set of thumb and index finger has the largest distance, so we can determine the index finger. Third, the little finger is determined as the farthest finger away from the thumb, and then the

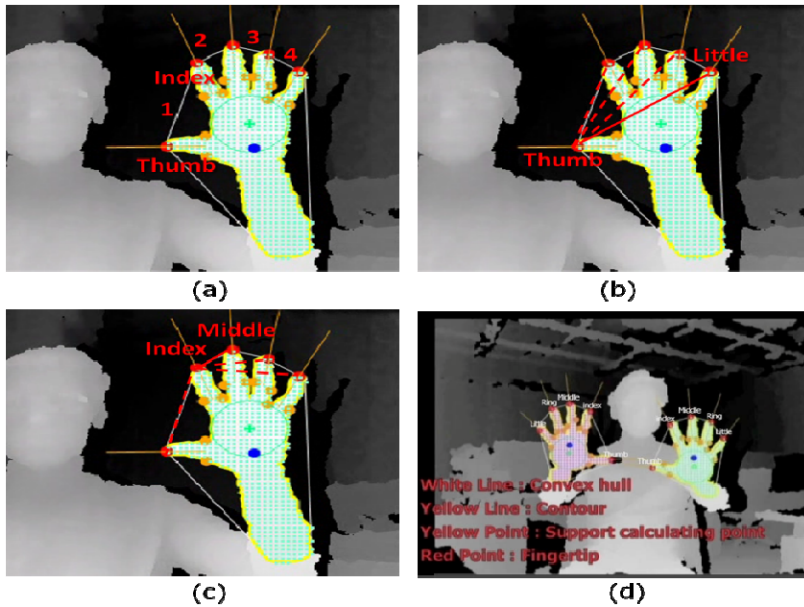


Fig. 3. Finger Identification. (a) Thumb and Index finger detection (b) Little finger detection (c) Middle finger detection (d) Identify fingers of two hands.

middle finger is determined as the closest finger from index finger (see (b) and (c) of figure 3). The remaining finger is determined as ring finger. The same method is used for identifying the fingers of both hands (see figure 3(d)).

4 Natural User Interface and Interaction

In this section, we explain the natural user interaction in each interface such as air-drawing, image manipulation and video manipulation. The interaction recognition method and feedback from the interface are explained in detail.

4.1 Air-Drawing Interface and Interaction

Our air-drawing interface implemented correctly finger painting interactions, by using depth-based finger gestures. We designed painting, drawing line gesture for these interactions. First, painting gesture is recognized when user extends his/her index finger, then the line is drawn along the path of movement.

Second, drawing line gesture is recognized when user extends their thumb and index finger, then the line is drawn depending on the coordinates value of the thumb and index finger (see figure 4(a)). The same interactions are made in the case of using both hands at simultaneously.

4.2 Image Manipulation Interface and Interaction

The Image manipulation interface is implemented well mapped with multi-touch interactions. It provides resizing (i.e. pinch, spread gesture) and rotating natural interaction with images.

In order to resize the image, the system tracks the number of fingertips. When system detects two fingertips (i.e. thumb and index finger of one hand or index finger of left and right hand), it changes to resizing interaction mode (see (c), (d) of figure 4). In resizing interaction mode, the system computes the distance between two fingertips. If the distance between fingertips becomes larger than a certain value found experimentally, zoom-in interaction will be performed and the size of the image is expanded. On the contrary, if the distance between fingertips is shorter than a certain value, zoom-out interaction will be performed and the size of the image is reduced.

In order to rotate the image, the system computes each fingertip's position (i.e. thumb and index finger coordinates of one hand or index finger coordinates for left and right hand). If the left index finger is raised upwards and the right index finger is moved downward then the image is rotated clockwise. On the contrary, if the right index finger is raised upwards and the left index finger is moved downwards, then the image is rotated counter clockwise (see figure 4(f)). When using only one hand, similar interactions are performed. If the thumb is raised upwards and the index finger is moved downwards, then the rotating interaction is performed in clockwise (see figure 4(e)).

4.3 Video Manipulation Interface and Interaction

Video manipulation interface implements interactions with video surface on 3D space using depth data. We designed the gestures for mapping, selecting surface and time shift interaction with media player.

Mapping interactions are made when system detects two fingertips on each hand. The video surface to play the video is created depending on the coordinates of the two fingertips of each hand (i.e. 4 sets of coordinates, thumb and index finger coordinate value for each hand). User can control the size of surface and mapping position by moving four points of P1,P2,P3 and P4 (see figure 4(b)). The surface's depth value can be controlled by moving a hand or both hands forward/backward. We can create a new surface in front of the surface that was created before, when both hands are moved forward. On the contrary, the new surface can be created behind the surface that was created before, when both hands are moved backward. In the case of moving a hand forward and the other hand backward, an almost diamond-shaped surface will be created. If shown from the front, it appears as a diamond surface. However, it can be shown as a rectangle surface when it is shown from different angles.

Selecting interactions occur when an index fingertip is on the surface that the user wants to control (i.e. it means that the depth value of index fingertip and the surface is same). This interaction is needed when multiple surfaces are created. The selected surface is shown with a white color border. Other interactions can be performed with the selected surface.

Time shifting interaction with selected surface occurs when the system detects five fingertips of the right hand and one fingertip of the left hand. The user can control the movie time bar by moving his/her left fingertip.

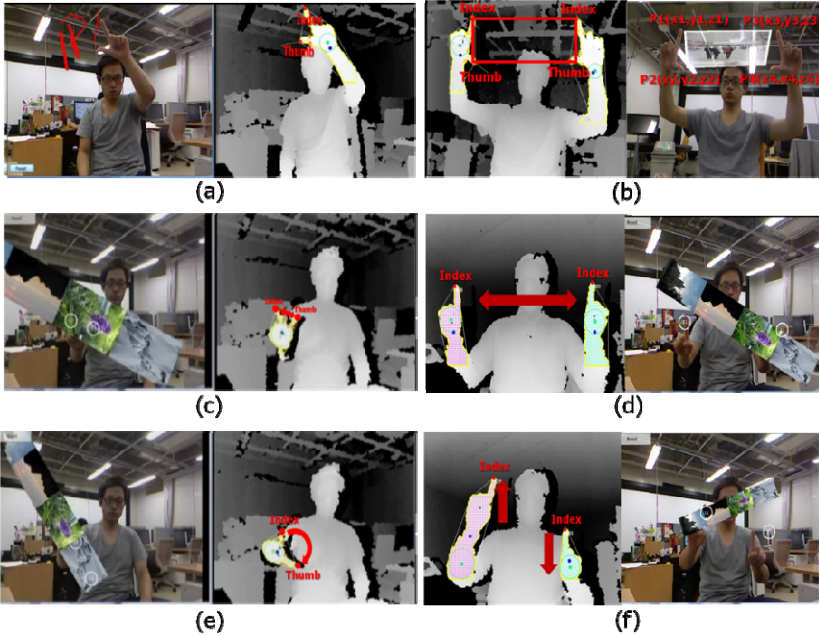


Fig. 4. Natural Interactions using Finger Gestures. (a) Drawing Gesture and Interaction (b) Mapping Gesture and Interaction (c) Spread Gesture and Zoom-in Interaction using a hand (d) Spread Gesture and Zoom-in Interaction using both hands (e) Clockwise Rotate Gesture and Rotating Interaction using a hand (f) Clockwise Rotate Gesture and Rotating Interaction using both hand.

5 Evaluation

5.1 Recognition Accuracy Experiment

In this experiment, we evaluated recognition accuracy with our proposed method for finger identification and designed finger gesture. The experiments were performed on a computer with Intel Core i5 CPU 2.67GHz and 4.0 GB RAM, using Microsoft Kinect for Xbox 360.

We performed the experiments with ten volunteers. Our experiments are designed to evaluate six gestures, i.e. extending all fingers for finger identification, drawing, pinch, spread, rotate and mapping gesture. After thoroughly explaining all our gestures, each volunteer performed a gesture 100 times, in each condition. We checked whether the system recognized the gesture or not. Figure 5 shows the average of recognition for the ten volunteers.

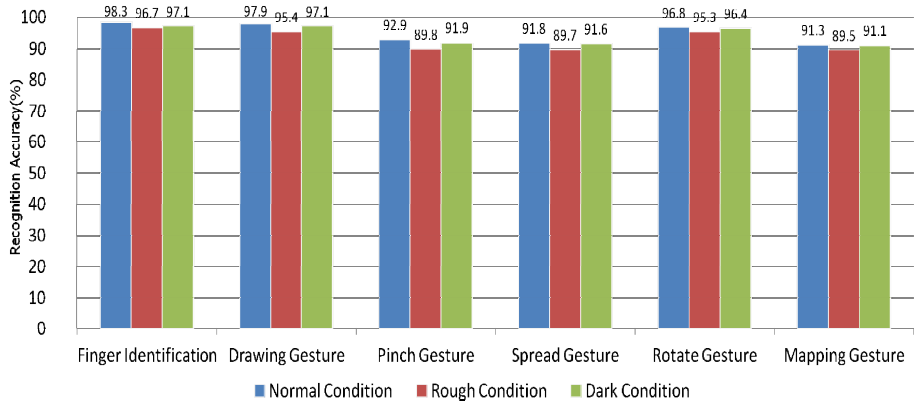


Fig. 5. The average result for ten volunteers

5.2 User Questionnaire

25 users participated in our system evaluation. They were all university students with various majors, 15 men and 10 women. They had experience in using Wii controller. We explained in simple terms our proposed interface's functions. The participants performed the proposed gestures and interactions freely, without any time limit. After completing their task, we asked them to answer a short questionnaire and offer a score from 1 to 5 (5 meaning very good) and comments.

Table 1. Average of 25 user's satisfaction for finger controller

Question	Average
Was it easy to interact using proposed gesture in each interface?	4.6
Was it natural to perform interaction and gesture using one hand?	4.64
Was it natural to perform interaction and gesture using both hands?	4.4
Was using the interface intuitive?	4.88
Was it natural compared with Nintendo Wii, Remote Game Conroller, Glove-based, etc?	4.92

5.3 Discussion

Our proposed method for finger identification and finger gesture recognition with depth data has shown high accuracy over 89 percentages for all gestures. Figure 5 shows almost the same accuracy with normal and dark conditions, and relatively low accuracy in rough conditions (e.g. many obstacles in background). However, the system was not influenced at all by dark and rough condition. We found that pinch,

spread and mapping gesture had a relatively low recognition rate. The pinch gesture was recognized as spread gesture because the users moved their hand rapidly, before the pinch gesture was recognized. The spread gesture was recognized as pinch gesture in the similar way. In addition, the mapping gesture has a high time complexity compared to other gestures, because the system calculates depth data from fingertips on both hands. Therefore, the recognition accuracy is decreased in case the user moves their hand rapidly, but it can be improved if we upgrade our system hardware.

Table 1 shows that our proposed interface and interaction are natural and intuitive. The final question had the highest score: it showed that our proposed wear-less interaction provides a more natural usage and intuitive feeling than wii, glove-based interface, and remote game controller. We received comments from the users about the need for applying this interaction in more practical situations, such as games, Google Street View, Google Earth. The users also suggested the use of our proposed interactions on larger displays.

6 Conclusions and Future Work

In this paper, we proposed a new natural user interaction using depth-based finger gestures. We presented finger tracking and finger gesture recognition techniques using finger identification with depth data. We use the clustering algorithm for hand detection and Graham Scan/Tracing algorithm for fingertip tracking. We propose our own method to identify fingers. After identification, the system detects the designed finger gesture from users and feedback. Our system provides gesture version of natural finger interaction and gestures such as drawing, resizing, rotating and mapping gesture. We presented the interfaces using our proposed interaction. These interfaces are : air-drawing, image manipulation and video manipulation.

We performed experiments for evaluating gesture recognition performance and we obtained satisfying results. They showed that finger gesture is promising for natural and complicate gesture recognition with high accuracy, i.e. over 89 percentages for all gestures. Our proposed interactions are natural and also intuitive.

In future work, we intend to support more practical functionalities with interaction. We also want to implement a 3d interface using HMD with finger gesture. It will need to apply object tracking for marker-less interface. We expect that this interface can include practical functions and will be more natural to use.

References

1. Ren, Z., Meng, J., Zhang, Z.: Robust Hand Gesture Recognition with Kinect Sensor. In: Proceedings of the 19th ACM International Conference on Multimedia, MM 2011 (2011)
2. Wigdor, D., Wixon, D.: Brave Nui World, pp. 9–15. Morgan Kaufmann (2011)
3. Cheoljong, Y., Yujeong, J., Jounghoon, B., David, H., Hanseok, K.: Gesture recognition using depth-based hand tracking for contactless controller application. In: 2012 IEEE International Conference on Consumer Electronics (ICCE), pp. 297–298 (2012)

4. Raheja, J., Ankit, C., Singal, S.: Tracking of fingertips and centers of palm using KINECT. In: 2011 Third International Conference on Computational Intelligence Modelling Simulation, pp. 248–252 (2011)
5. Chu, S., Tanaka, J.: Hand Gesture for Taking Self Portrait. In: Jacko, J.A. (ed.) Human-Computer Interaction, Part II, HCII 2011. LNCS, vol. 6762, pp. 238–247. Springer, Heidelberg (2011)
6. Kenn, H., Megan, F., Sugar, R.: A glove-based gesture interface for wearable computing applications. In: Proceedings of the IFAWC 4th International Forum on Applied Wearable Computing 2007, pp. 169–177 (2007)
7. Lenman, S., Bretzner, L., Thuresson, B.: Computer Vision Based Hand Gesture Interfaces for Human-Computer Interaction. Technical Report CID-172, Center for User Oriented IT Design, pp. 3–4 (2002)
8. Wachs, J., Kölsch, M., Stern, H., Edan, Y.: Vision-based hand-gesture applications. *Communications of the ACM* 54(2), 60–70 (2011)
9. Frati, V., Prattichizzo, D.: Using Kinect for hand tracking and rendering in wearable haptics. In: IEEE World Haptics Conference, pp. 317–321 (2011)
10. Graham, R.: An efficient algorithm for determining the convex hull of a finite planar set. *Information Processing Letter*, 132–133 (1972)
11. Iwai, Y., Watanabe, K., Yagi, Y., Yachida, M.: Gesture recognition using colored gloves. In: IEEE Int. Conf. Pattern Recognition, vol. A, pp. 662–666 (1996)
12. Nanda, H., Fujimura, K.: Visual tracking using depth data. In: Conference on Computer Vision and Pattern Recognition Workshop, p. 37 (2004)
13. OpenNI organization. OpenNI User-Guide (2012)
14. He, G., Kang, S., Song, W., Jung, S.: Real time gesture recognition using 3D depth camera. In: 2011 IEEE 2nd International Conference on Software Engineering and Service Science (ICSESS), pp. 187–190 (2011)
15. Tang, M.: Hand Gesture Recognition Using Microsoft's Kinect. Paper Written for CS228, Winter 2010. Technologies, UIST 2011, pp. 1–9. ACM (2011)

A Method for Single Hand Fist Gesture Input to Enhance Human Computer Interaction

Tao Ma¹, William Wee¹, Chia Yung Han², and Xuefu Zhou¹

¹ School of Electronic and Computing Systems, University of Cincinnati, USA

² School of Computing Sciences and Informatics, University of Cincinnati, USA
mata@mail.uc.edu, {weewg, han, zhoxu}@ucmail.uc.edu

Abstract. The study of detecting and tracking hand gestures in general has been widely explored, yet the focus on fist gesture in particular has been neglected. Methods for processing fist gesture would allow more natural user experience in human-machine interaction (HMI), however, it requires a deeper understanding of fist kinematics. For the purpose of achieving grasping-moving-rotating activity with single hand (SH-GMR), the extraction of fist rotation is necessary. In this paper, a feature-based Fist Rotation Detector (FRD) is proposed to bring more flexibility to interactions with hand manipulation in the virtual world. By comparing to other candidate methods, edge-based methods are shown to be a proper way to tackle the detection. We find a set of "fist lines" that can be easily extracted and be used steadily to determine the fist rotation. The proposed FRD is described in details as a two-step approach: fist shape segmentation and fist rotation angle retrieving process. A comparison with manually measured ground truth data shows that the method is robust and accurate. A virtual reality application using hand gesture control with the FRD shows that the hand gesture interaction is enhanced by the SH-GMR.

1 Introduction

Hand gesture recognition is a mathematization of the interpolation of human hand gestures assisted by modern computer technology. The purpose is to replace traditional input devices, keyboard and mouse, with a new fashion that makes human interact with computer in a way that is as natural as in the real world [1]. In spatial domain, static hand gestures are recognized due to the different spatial distribution of fingers with respect to palms. A "thumb-up" gesture in sign language means "good". A "fist" gesture might stands for "stop" [2] or other meanings [3]. Laura et al presented a joint segmentation and adaptive classifier that can discriminate 4 static hand gestures under slight occlusion condition [4]. Yi et al.'s classifier that combined both supervised and unsupervised training process recognizes 14 hand gestures [5]. On the other hand, in temporal domain, hand kinematics under various gestures is meaningful as well. Yi and Thomas [6] described articulated hand local motion for a 16 rigid object 3D hand model with inverse kinematics. Human hand motions are very complicated and always occur in both spatial and temporal domain. A well representation of hand

gesture patterns and their kinematics lies in the improvement of natural user experience-as if users interact with the real world. However, current gesture recognition methods can hardly capture all subtle movements for manipulation in the virtual world.

Holding an object and moving it with single hand is a common activity in daily life. Clenching fingers together to form a fist is a natural gesture, which represents grasping [7]. The ability to map this activity in designing interactions that allow inputs for software applications can be very useful. However, the processing of images of hand fist and characterization of the various motions as input to computer is not straightforward. We term this motion behavior as single hand grasping-moving-rotating (SH-GMR). These three actions always occur simultaneously. The detection of moving, or simple translation, is fairly easy, attested by the fact that a variety of algorithms already exist [8]. However, the detecting and tracking on fist rotation has been lacking for a long time. The existing compromised solution of rotating a virtual object is to make use of two hands to decide the rotation angle, which includes rotation about an axis and "steering wheel" rotation [9]. The drawbacks are obvious. Two hands have to grasp the same object at the same time. For tiny objects, it needs supplementary visual guidance for users to hold, such as virtual handles attached on the objects. More importantly, this two-hand gesture prevents users to interact with two objects at a time, which makes the interaction manners of many kinds quite awkward and inefficient. Thus, user experience suffers greatly. With these concerns, we argue that the single hand rotation is better than the two-hand rotation. Moreover, the study of fist rotation is crucial to achieve SH-GMR.

From the image and vision perspective, the fist rotation is defined as a 3 dimensional (yaw, pitch and roll) rotations of a deformable, scale variable and intensity variable object with associated translation movements. It is necessary to distinguish fist from other gestures because other gestures extend at least one finger out [10]. But fist gesture is defined as a hand with all fingers clenched into the palm [11]. Within certain angle of view, fingers are still visible, but they are fully folded and placed side by side with each other. In the following sections, we discuss several methods that can potentially be used for the fist rotation extraction and then give a proper solution to tackle this problem.

2 Related Research

There are several methods that could potentially be used as fundamental methods of FRD. They are skeletal model, volume model, optical flow, local features, and edge features. We review them below and give examples if necessary.

Current skeletal models for kinematic representation of hand are applied for the open handed gesture that at least one finger is stretched out. Lee and Kunii [12] introduced a 27 DOFs hierarchical skeletal hand model with constraints on joint movements, which makes some hand configurations impossible so as to reduce the shape ambiguity. Du and Charbon proposed a 30 DOFs skeletal model for depth image fitting [13]. These models are suitable for describing the clenched fist, but they

did not propose how to apply the model fitting algorithms in extracting the fist rotation using their internal constraints and external image forces.

3D dense map, sometimes called point cloud generated by structured light [14], TOF camera or stereo camera, can be used to solve the ambiguity in certain degree. But huge computational cost of 3D fitting prevents them from being applied in real-time tracking [15]. We have not seen any solution that can handle the fist rotation problem. Besides, the depth map captured by current cameras cannot provide enough details in resolution to extract individual fingers from a clenched fist.

Due to the complexity of hand movements, local features should be found that are invariant to hand translation, scaling, rotation, and invariant to illumination changes and 3D projections. We applied the Scale Invariant Feature Transform (SIFT) method [16, 17] to fist rotation detection to see whether it works properly. We test the hand feature matching problem using SIFT demo program developed by Lowe. Our goal is to see whether SIFT can constantly track the same features in continuous frames as long as they are visible. The matching features are connected by straight lines in azure in Figure 1. In each image pair in Figure 1, hand on left side is always the 1st frame of the video; right hands are in the 2nd, 5th, 15th, 30th, 45th, and 65th frames respectively. SIFT finds most of feature pairs correctly, but the number of points reduces sharply as the angle difference becomes larger. The method casts away most of detected features to keep a correct matching, which is not suitable for the extraction.

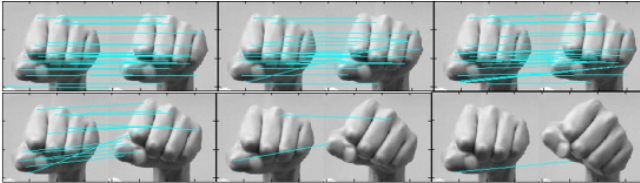
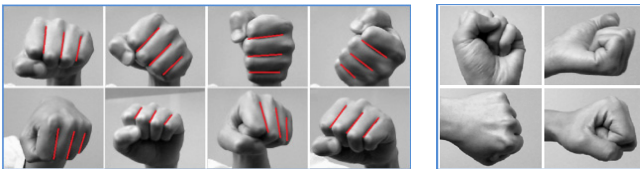


Fig. 1. SIFT matching under different fist angles, between 1st frame and 2nd, 5th, 15th, 30th, 45th, 65th frames



(a)

(b)

Fig. 2. Edge features on fists. (a) Manually labeled fist lines under different rotations. (b) Hand gestures that do not fully show the fist lines.

It is easy to find that when five fingers are clenched, brightness between two fingers is darker. Between index, middle, ring and pinky fingers, 3 clearly dark lines can be seen under most lighting conditions. These lines are nearly straight, parallel, and almost appear or disappear at the same time. During the rotation, they maintain their

relative positions unchanged no matter whether fist is facing directly to the camera or not. They are also good to preserve the fist structure. They are more stable and accurate to be tracked, compared with methods listed above. Moreover, instead of extracting relative angle value from frame-to-frame computation, the angle of the lines are absolute value with respect to camera coordinates, which means no cumulative error would be established. We call these 3 lines as "fist lines" for simplicity. As shown in Figure 2(a), fist lines are manually labeled as red color in various fist rotations. The fist lines will not appear under certain field of views. Figure 2(b) shows some fist gestures in which the fist lines are not clear or cannot be found. So far, the edge features turn out to be the best approach to handle the fist rotation.

3 Approach

In this paper, it is assumed that human arms have been segmented from the whole images, and also hands are in fist shape. Arm segmentation from other parts of the human body and fist shape classification from other gestures are beyond the scope of our interests.

3.1 Fist Shape Segmentation

Observation shows that, if seeing along a human arm, the width of the fist is always larger than that of the forearm under all camera views. Values of the width would have a suddenly drop down if sliding from fist side to arm side. This geometrical characteristic is feasible for the fist segmentation. Arm shapes in 2D are first transformed to 1D representation through a dimension reduction process, and then a classifier is used to decide the fist position along the arm.

A contour retrieving algorithm is applied to topologically extract all possible contours in the image [18]. *Contour C* with the largest number of point set is the outermost contour of the arm, shown as Figure 3 (a). Using the data set of the *contour C*, a convex hull and its vertex set *P* [19] are computed. Sometimes image noise causes trivial boundary so that the number of vertices is sharply increased. In this case, a polygon approximation routine is used to reduce the excessive details along the boundary. The number of vertex should better be in the range of 8 to 15 considering both computational cost and accuracy. We compute the Euler distances of all vertex pairs except those who are adjacent. Then we find the longest two distances \vec{a} and \vec{b} . The direction of the main axis *l* is set to be the bisector of the angle of the two vectors:

$$\theta_{ma} = \frac{1}{2} a \cos\left(\frac{\vec{a} \cdot \vec{b}}{|\vec{a}| |\vec{b}|}\right) \quad (1)$$

Two longest distances are used to decide the main axis to prevent the value oscillation introduced by noise. Then, the whole points in *contour C* are rotated and translated:

$$contourC = \begin{bmatrix} \cos \theta_{ma} & -\sin \theta_{ma} \\ \sin \theta_{ma} & \cos \theta_{ma} \end{bmatrix} contourC + T \quad (2)$$

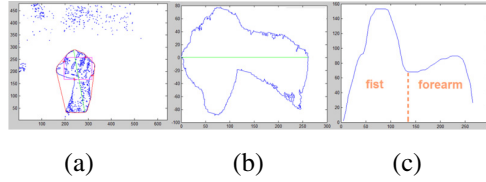


Fig. 3. Fist segmentation. (a) the arm contour, convex hull, main axis, and the search window marked on the contour image, (b) a fist contour that has been rotated and translated to horizontal position, (c) width curve of the contour in (b).

The rotated main axis l' is located on x-coordinate, shown in Figure 3 (b). Given a x on the rotated contour, two y values are corresponded: one negative and one positive. The difference of these two y values indicates the width of the fist along l' . Computing all the widths along l' , we get a smoothed width curve of the contour C' , as shown in Figure 3(c). So far, we convert a 2D shape clustering problem to 1D. It is easy to classify fists from forearms by looking at features along the width curve. Various clustering methods can be used, such as K-means, area-based, and etc. The curve on Figure 3 (c) shows that the fist is located at the first half of the contour of Figure 3 (b). Going back to the original contour C , a fist bounding box is found according to the result of the width curve, shown in Figure 3 (a) in a magenta rectangle. The bounding box is served as the search window for the fist detection.

3.2 Fist Rotation Detection

Finding the three fist lines is a challenging task for the reason that there are many line and curve features in the search window. Inspired by the observation shown in Figure 2, we find the fist lines are basically straight, parallel, and almost appear or disappear at the same time. Thus, three parallel straight lines with the interval of d are used as the theoretical model to fit the selected feature point data. 3 parameters need to be decided: the slope of the lines θ , the intercept of the middle line b , and the interval d . Note that even though the fist lines are equidistant, their distances appearing on images may not be the same due to the perspective from 3D space to the camera plane. But in this paper, we particularly see the roll rotation as the major direction meanwhile ignoring other DOFs. The mathematical model is:

$$\begin{cases} y = x \tan \theta + b + d / \cos \theta; \\ y = x \tan \theta + b; \\ y = x \tan \theta + b - d / \cos \theta \end{cases} \quad \text{where } -90 < \theta < 90 \quad (3)$$

Edge Feature Extraction. We use Laplacian of Gaussian (LOG) [20] method to extract features in the search window because it is scale sensitive to blobs that has the similar size. It has strong response to features of extent $\sqrt{2\sigma}$, where σ is the variance of Gaussian function. The LOG kernel can be pre-computed before the convolution on the original image:

$$\nabla^2 G(x, y) = -\frac{1}{\pi\sigma^4} \left[1 - \frac{x^2 + y^2}{2\sigma^2} \right] e^{-\frac{x^2 + y^2}{2\sigma^2}} \tag{4}$$

Since features around fist lines are very similar, the LOG method extracts stable and mostly continuous lines between fingers. But it also recovers features located on hand edges. Figure 4 shows edge feature maps under different fist rotations. All the edges are stored into a structured 2-row array E . Mathematically, this edge set E is $E = \bigcup_N \varepsilon$, where N is the number of feature edges ε .

A Rough Angle Estimation with Histogram. The edge feature maps gives us an important clue that after the feature extraction the amount of features on the fist lines is larger than that on other non-fist-line features. It is because the fist segmentation preserves most of fist lines while eliminating most of unrelated features. Figure 4 shows that the distribution of the histograms of the feature maps is highly related to the fist angles. We can approximately compute the coarse fist angle range δ by finding the highest percentage of pixel bins within the histogram of the slopes. The highest bins and its two nearest neighbors are picked to calculate the angle range using the center of gravity method. To compute the slope of the feature segments, the step length between two points should be larger than 5 so as to provide plenty of angle resolution.

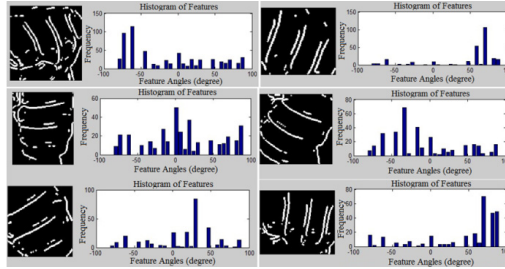


Fig. 4. Feature maps and their corresponding histogram under different angles

Back Projection and Edge Pruning. All pixels within the angle range δ are back-projected to the edge feature map. Edges that contain these pixels are marked and kept in Ω , while other edges are pruned:

$$\Omega = \bigcup_{\varepsilon \in E} [\varepsilon \mid \exists \gamma_\varepsilon : \gamma_\varepsilon \in \delta] \tag{5}$$

where γ_ε is the angle value of the feature segments in edge ε . After this process, the number of edge candidates is greatly reduced.

Cutting off, Merging, and Sorting Operation. In Ω , the slope of features within one edge may go out of the angle range. These parts are cut off from the edge and the residues are merged again, indicated as:

$$\varepsilon' = \bigcup_{\chi_\varepsilon \in \varepsilon} [\chi_\varepsilon \mid \gamma(\chi_\varepsilon) \in \delta] \tag{6}$$

where χ_ε is a feature point in ε , and $\gamma(\chi_\varepsilon)$ is the angle value of χ_ε .

If two edges are almost collinear as if they are in the same fist line, they are merged into one edge, described as:

$$\varepsilon_{12} = \bigcup_{\substack{\chi_{\varepsilon_1} \in \varepsilon_1 \\ \chi_{\varepsilon_2} \in \varepsilon_2}} [\chi_{\varepsilon_1}, \chi_{\varepsilon_2} \mid \forall \chi_{\varepsilon_1}, \forall \chi_{\varepsilon_2} : \text{collinear}(\chi_{\varepsilon_1}, \chi_{\varepsilon_2})] \tag{7}$$

where ε_{12} is the merged feature set from ε_1 and ε_2 , and $\text{collinear}(\)$ is a function that decide whether two feature points are basically collinear.

Last, all the existing edges are sorted according to their positions, and then stored in set Y . So far, the number of edges in Y is slightly more than 3, which is very cost effective for the following angle refining process.

Fitting the Mathematical Model with the 3 Selected Edges. For any given 3 edges, $\varphi_1, \varphi_2, \varphi_3$ in Y , parameters θ, b , and d can be calculated by fitting the theoretical model described in equation (3) to the three edges. To convert equation (3) into linear equations, we let $k = \tan\theta$, and $c = d/\cos\theta$. Then the equations can be expressed with linear equations as:

$$A \tilde{x} = \tilde{y}, \quad \text{where}$$

$$A = [k, b, c], \quad \tilde{x} = \begin{bmatrix} x_{11} & x_{12} & x_{21} & x_{22} & x_{31} & x_{32} \\ 1 & 1 & \dots & 1 & 1 & \dots & 1 & 1 & \dots \\ 1 & 1 & 0 & 0 & -1 & -1 & \dots & \dots & \dots \end{bmatrix}$$

$$\tilde{y} = [y_{11}, y_{12}, \dots, y_{21}, y_{22}, \dots, y_{31}, y_{32}, \dots],$$

$$(x_{1i}, y_{1i}) \in \varphi_1, (x_{2i}, y_{2i}) \in \varphi_2, (x_{3i}, y_{3i}) \in \varphi_3. \tag{8}$$

Several methods can be used to solve this over-determined, multiple linear regression problem, such as least square, Gauss elimination, and Singular value decomposition (SVD). The fitting error $E(\varphi_1, \varphi_2, \varphi_3)$ can be derived from the sum of absolute difference (SAD) between fitted lines and edge pixels:

$$E(\varphi_1, \varphi_2, \varphi_3) = \sum_{j=1}^3 \sum_{(x_j, y_j) \in \varphi_j} \frac{|kx_j - y_j + b + (2-j)c|}{\sqrt{k^2 + 1}} \tag{9}$$

Optimized Fist Lines with Minimum Error. We compute all the combinations of 3 possible fist lines in Y . The number of combination is given by C_n^3 , where n is number of edges in Y . A correct choice of the fist lines is indicated by $(\varphi_1^*, \varphi_2^*, \varphi_3^*)$ that has the minimum fitting error:

$$(\varphi_1^*, \varphi_2^*, \varphi_3^*) = \arg \min_{\varphi_1, \varphi_2, \varphi_3 \in Y} E(\varphi_1, \varphi_2, \varphi_3) \tag{10}$$

Its angle $\theta^* = \text{atan}(k^*)$ is the optimized fist rotation angle within $(-90^\circ, 90^\circ)$. A large amount of pixel involved in the fitting process guarantees an accurate and stable outcome. Figure 5(a) shows the three fitting lines are found, marked with red, blue, and magenta.

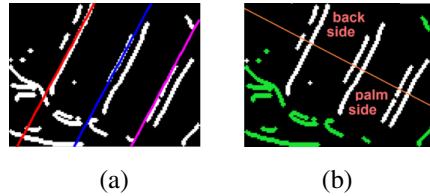


Fig. 5. (a) Three fitting lines that minimum the fitting error. (b) Feature points out of the angle range marked with green color.

Deciding the Fist Rotation within 360° . As mentioned above, the line model can only decide rotation within $(-90^\circ, 90^\circ)$. Due to the special finger position with relation to the palm, more features can be found near the palm side rather than the back side of the fist. These features mostly have different directions with the fist lines. We empirically discriminate between the palm side and the back side by measuring the distribution difference of features that are out of the range of the rough rotation angle δ . They are marked as green color in Figure 5(b). The center of gravity of the selected 3 fist lines are first computed. Then, through this point, a straight line (the orange line in Figure 5(b)) that is perpendicular to the fist lines splits the search window into two parts. The palm side and back side of hands must be located in these two parts respectively. The part that has more green pixels is the palm side, and vice versa. With the angle θ^* computed in the previous steps, the final rotation angle can be decided within 360° .

4 Experiment and Application

We pay mostly attention to the accuracy and stability of the proposed FRD. One experiment is implemented to test these two aspects. To generate ground truth (GT) data, two markers in cross shape are stuck on the middle finger so that they can be manually labeled afterwards and be used to calculate hand rotations, shown as Figure 6. Then the GT angle is compared with the angle generated by the FRD every 10 frames.

The GT angle is manually computed every 10 frames. Then it is compared with the FRD output within the same frame, shown as Figure 7. The maximum angle value is 140° due to the physical limit of human hands. The result shows that the proposed FRD method is stable and consistent with the GT data, with the absolute mean difference of 3.27° (0.9% of 360°), and standard deviation of 2.81° (0.8% of 360°). The largest error occurs between 110° and 140° .

There are several reasons that cause the error. First, the manually labeled GT value may not be accurate due to the image quality. Then, remember that hand is a deformable object. The rotation of the markers may not fully represent that of the fist lines, especially when the hand is twisted almost to its physical limit. Last, as analyzed in previous sections, hand rotation always happens in 3 DOFs. The proposed FRD and the GT measurement only consider one major movement while ignoring others. This will also introduce difference in the comparison.

Considering the real applications of fist rotations in HMI and the accuracy level of human body movements, the proposed FRD is effective to be used as detecting human fist rotations in HMI applications. The computational cost is various depending on the amount of feature points. In our test video with the resolution of 640×480 , the size of the search window after the fist shape segmentation is usually within 120×120 . Modern computer can easily handle this amount of data in real time.

To illustrate the improvement of the SH-GMR behaviors with the help of the proposed FRD algorithm, we present a simple application that implements a chemical reaction experiment in a 3D virtual reality environment. This system captures 3D hand movements with stereo cameras. With our FRD routine integrated, the system is able to handle SH-GMR. Figure 8 shows a user is implementing a chemical experiment by pouring one sort of liquid from the right flask into the left flask to trigger certain chemical reaction. As shown in Figure 8 (b), in the virtual environment, two flasks are moved close to each other, and right flask is leant to pour the liquid into the left one, with the operation of two hands respectively.



Fig. 6. Two markers stuck on a hand for computing GT angle.

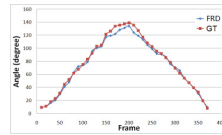
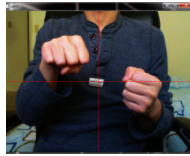
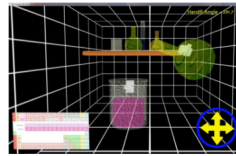


Fig. 7. The comparison of the FRD and GT for every 10 frames



(a)



(b)

Fig. 8. (a) The user's gesture of operating objects in the virtual reality environment. (b) An ongoing virtual chemical experiment controlled by two hands.

5 Conclusion

Hand interaction is highly limited by the current two-hand rotation gesture due to the lack of the research on hand fist kinematics. A single fist rotation detector is crucial to implement single hand grasping-moving-rotating activity that makes two hands fully control different objects possible. We present a feature-based FRD method that provides accurate and stable detection of the fist rotation problem for the purpose of enriching hand gesture databases with finer hand motion sequences. Except the fist rotation, there are plenty of hand gestures and their kinematics that we have not fully utilized. Deeply digging into this area will greatly benefit the hand gesture interaction and also bring user experience to a brand new level.

References

1. Beale, R., Edwards, A.D.N.: Gestures and Neural Networks in Human-computer Interaction. In: IEE Colloq. Neural Nets in HCI (1990)
2. Manresa, C., Varona, J., Mas, R., Perales, F.J.: Real-Time Hand Tracking and Gesture Recognition for Human-Computer Interaction. In: ELCVIA (2000)
3. Binh, N.D., Shuichi, E., Ejima, T.: Real-Time Hand Tracking and Gesture Recognition System. In: GVIP (2005)
4. Gui, L., Thiran, J.P., Paragios, N.: Joint Object Segmentation and Behavior Classification in Image Sequences. In: CVPR (2007)
5. Wu, Y., Huang, T.S.: View-independent Recognition of Hand Postures. In: CVPR, vol. 2, pp. 88–94 (2000)
6. Wu, Y., Huang, T.S.: Capturing Articulated Human Hand motion: A Divide-and-conquer Approach. In: ICCV, vol. 1, pp. 606–611 (1999)
7. Hooker, D.: The Origin of the Grasping Movement in Man. In: Proceedings of APS, vol. 79(4), pp. 597–606 (1938)
8. Bradski, G., Kaehler, A.: Learning OpenCV, 2nd edn. O'Reilly Media (2008)
9. Hinckley, K., Pausch, R., Proffitt, D., Kassell, N.F.: Two-Handed Virtual Manipulation. In ACM Trans. CHI 5(3), 260–302 (1998)
10. Triesch, J., Malsburg, C.V.D.: A System for Person-Independent Hand Posture Recognition against Complex Backgrounds. IEEE Trans. PAMI 23(12), 1449–1453 (2001)
11. Kingston, B.: Understanding Joints: A Practical Guide to Their Structure and Function, 2nd edn. Nelson Thornes (2000)
12. Du, H., Charbon, E.: 3D Hand Model Fitting for Virtual Keyboard System. Applications of Computer Vision. In: IEEE WACV (2007)
13. Rosales, R., Athitsos, V., Sigal, L., Sclaroff, S.: 3D Hand Pose Reconstruction Using Specialized Mappings. In: ICCV, vol. 1(200), pp. 378–385 (2001)
14. Kollarz, E., Penne, J., Hornegger, J., Barke, A.: Hand Gesture Recognition with A Novel IR Time-of-Flight Range Camera-A pilot study. Int. Journal of Intelligent Systems Technologies and Applications Archive 5(3/4), 334–343 (2008)
15. Bruce, L.D., Takeo, K.: An Iterative Image Registration Technique With An Application to Stereo Vision. Proceeding IJCAI 2, 6474–6679 (1981)
16. Lowe, D.G.: Object Recognition from Local Scale-invariant Features. In: IEEE Conf. CV, vol. 2, pp. 150–1157 (1999)
17. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. Int. Journal of Intelligent Systems Technologies and Applications Archive 60(2), 91–110 (2004)
18. Homma, K., Takenaka, E.I.: An Image Processing Method for Feature Extraction of Space-occupying Lesions. JNM 26, 1472–1477 (1985)
19. Marr, D., Hildreth, E.: Theory of Edge Detection. Defense Technical Information Center (1979)
20. Suzuki, S.: Topological Structural Analysis of Digitized Binary Images by Border Following. ACM CVGIP 30(1), 32–46 (1985)

Kinect© as Interaction Device with a Tiled Display

Amilcar Meneses Viveros¹ and Erika Hernández Rubio²

¹ Departamento de Computación, CINVESTAV-IPN, México D.F.

`ameneses@cs.cinvestav.mx`

² SEPI-ESCOM, Instituto Politécnico Nacional, México D.F.

`ehernandezru@ipn.mx`

Abstract. The use of high resolution tiled display has become popular in the scientific community. User interaction with these devices depends on the hardware configuration and the software in use. The variety of hardware configurations and software generates various types of execution modes and interaction in the tiled display, this diversity has resulted in not having a standard for human computer interaction. This paper shows the results of the interaction between users and the tiled display using the Kinect©. The results help us find improvements in hardware configurations of this arrays of displays, applications design and try to find standards in defining user-defined motion gestures.

1 Introduction

The use of high resolution tiled display has become popular in the scientific community [TS][SC][MC] [NH][NV][CC][ET]. They are used for collaborative work and for the deployment of information and navigation of large data volumes [SC][MC][NV][BG]. Current research on human computer interaction in the tiled display is focused on two main points: the perception of information displayed [ET][YN][BN1] [BN2][TGSP] and control applications [BN2][AT][AG][BG][NJ]. User interaction with these devices depends on the configuration of hardware for controlling video wall (visualization server or visualization cluster) and the applications running on it [TS][NH][CC][BN1] [RCMM]. Applications running on these devices can be of two types: distributed applications or desktop applications [TS][MH][NJ].

Distributed applications are developed to take advantage of the hardware features of the visualization cluster [PR][HH][GdS]. This applications are highly scalable. When the application is running on the cluster, it is very common to use the master node (or front-end node) as responsible for the interaction with the application running on the tiled display. These applications are for very specific purpose and user interaction with the device is made from the master node. In this case, the user interaction with the tiled display is limited to a set of basic operations that fulfill the functionality of the application [TS][NH][CC][AG][R][BL][NJ].

Desktop applications take advantage of the operating system's ability to export graphical interface to tiled display either the case of a visualization cluster or visualization server [TS][NJ]. The scalability of desktop applications depend on the capabilities of operating system and graphics cards. When wing applications running in an extended desktop, the user can work through the master node (for example, in case of using the VNC protocol) or directly on the arrays of displays (for example, when using XDMX). User interaction is with the window manager or the graphical user interface of the operating system. The user interacts with the window manager using mouse, keyboard or combination of both.

The variety of hardware configurations and software generates various types of execution modes and interaction in the tiled display, this diversity has resulted in not having a standard for human computer interaction[YN]. On the one hand the problem is show de information for the users and the other hand the application control. Furthermore, when adding new specific use applications, they are restricted to only user interaction that is in the master node [TS][CC][HH][NJ]. User interaction is limited to the capabilities of the master node. When applications working at a desktop, some interactions mechanism can have a critical impact, for example, mouse manipulation, can cause functionality problems if not cared Fitt's law [BN1][BN2]. It is possible to take advantage of MOCAP technology, specifically the Kinect©, to provide the user with an intuitive means that permits them to interact with tiled display applications. This approach allows user mobility along the tiled display freeing interaction with the master node and to define a set of gestures to control applications running on the tiled display.

In this paper, we present the results and proposals to use the Kinect©as interaction device with a tiled display. This tiled display is controlled by a cluster of Apple Mac Mini [MC]. Interaction experiments reported in this paper are related to the control of the tiled display desktop through VNC protocol and several special-purpose distributed applications, such as an image viewer, a web browser and ParaView. The applications have been selected to cover the widest possible use case. For interaction with applications, we tested a set of gestures (proposed in other works and added some) [AG][KB], that through Kinect©, are interpreted to perform specific actions on the tiled display or in the desktop of the master node. Tests were conducted to emulate the behavior of the mouse and keyboard. In particular, for the case of interaction with widgets that respond to the keyboard events, such as text fields,we use the technique of virtual keyboards (common in environments of smartphones with touch screens), and we did a comparative metrics based on usability, functionality and effectiveness between the keyboard virtual (stylus based) and 8pen Android keyboard.

2 MOCAP Technologies

The motion capture, motion tracking or MOCAP are terms used to describe the recording process and translating of that motion into a digital model. This

Technology has been used for capture and analysis of the human movement. Control applications where estimated motions can control something benefit about these technologies as interfaces with gaming, virtual reality and human computer interfaces. There are many MOCAP technologies, the main categories of these are using markers or without the markers to recognize parts of the human body. The Kinect is a technology that does not use markers for this purpose.

2.1 Kinect©

The Kinect© is a free gaming device developed by Microsoft© for the XBOX 360 and PC in the future through Windows 8. Kinect© enables users to control and interact with the console without having physical contact with a traditional video game control, by a natural user interface that recognizes gestures and voice commands.

The main objective of this device is increase the use of the Xbox 360. In our case we use this device to allow a user to control a tiled display, using only the movements and efforts of his body. Kinect© incorporates different technologies as an emissary of infrared that emits an invisible light which together with a CMOS sensor looks how this emission is reflected back and passes de data to the console or in this case to the PC in a grayscale format so it can determinate the depth of the scene and the motion of the user.

The algorithm which operates under the Kinect© was developed by Microsoft Research Centre in Cambridge. The algorithm takes as an input a depth image which is analyzed pixel by pixel, where each pixel is assessed according to their characteristics such as its depth, if the pixel belongs to an upper or lower part of the body; the result of each characteristic about pixel is combined with the research in a classifier called forest of decision thus a collection of decision trees where each decision tree was trained or specialized in a set of characteristics of an image depth which different parts of the body were previously labeled. The classifier assigns certain probability to the pixel belongs to a particular body part, then the algorithm assigns the pixel belonging to a body part that have obtained most probability in the classifier. Finally which each pixel belonging into a particular body part, hinge points are assigned to the identified areas in the body in three views, frontal, side and upper[MRCC]. This procedure is made 3 times per second; this is analyzed 200 pictures or images per second, which is about 10 times faster than other techniques for recognition and motion detection.

3 Experiment

We are interested in to know the interaction of the user with the tiled display through the Kinect©. HCI metrics that we are using are usability, performance, effectiveness and comfort. For this purpose we have developed several questionnaires that apply to users after an action or task running in the tiled display. We applied twelve questionnaires. Which includes actions such as moving the

-
1. How easy it was to control the mouse pointer through the Kinect?
a) Very Diffcult b) Difficult c) Easy d) Very Easy
 2. How fast did you control the mouse pointer through the Kinect?
a) Vey Slow b) Slow c) Fast d) Very Fast
 3. How comfortable did you control the mouse pointer through the Kinect?
a) Very Confortable b) Confortable c) Uncomfortable d) Very Uncomfortable
 4. How fast was the response of the mouse pointer using the Kinect?
a) Vey Slow b) Slow c) Fast d) Very Fast
 5. Did it seem appropriate at the time it took to answer the pointer?
a) Yes b) No
 6. Controlling mouse pointer through the Kinect, was successful?
a) Yes b) No
 7. The mouse pointer control, through the Kinect, is performed as you thought?
a) Yes b) No
-

Fig. 1. Example of questionnaires used in this experiment

cursor, select a character from virtual keyboard until tasks such as opening an application, close a window and delete text, to name a few. An example of this type of questionnaire is shown in the Figure 2.1.

We applied two types of tests. The first is to study the interaction of the user with the desktop and specific applications. The second is to compare the use of two types of virtual keyboards: Stylus based and Android 8pen. The tests were performed with two different groups of students: We use a sample of twenty people for the first test and twenty one for virtual keyboard tests. Ten questionnaires were used for the first test and two questionnaires for the second test.

3.1 User-Defined Motion Gestures for Tiled Display Interaction

User-defined motion gestures are associated to actions and task. A task is a sequence of actions. In general, the actions that we consider in this work are move de the mouse, push the button and select a character.

The gestures considered for the first test (use of applications and desktop) are:

- Move the mouse pointer** The gesture associated with this action, is to point the palm towards the Kinect ©, each movement (up, down, left or right) will have the effect to control the mouse pointer.
- Click** The gesture associated with this action, is to point the palm towards the Kinect©, but making a move forward and return to the original position.
- Item selection** This is done using the click action. Is to point the palm towards the Kinect ©, but making a move forward and return to the original position. Allows point and select objects on the screen. The system applies to a function or process these objects.
- Open Application** This gesture is similar to the Item Selection task. Is to point the palm towards the Kinect©, but making a move forward and return to the original position. Allows point and select objects on the screen. The system applies to a function or process these objects.
- Drag Item** It's like the click action, except that the hand stays on and does not return to its original position. Allows you to move (drag) an object on the screen or rotate it.
- Open Context Menu** The gesture associated with this task is to carry out click movement, with the helping hand (one that does not control the mouse pointer).
- Rotate Object** It's like the click action, except that the hand stays on and does not return to its original position. Allows you to move (drag) an object on the screen or rotate it.
- Slide Up/Down** This gesture is to put the palm facing the Kinect© and make a move down quickly.
- Slide Left/Right** Involves placing the palm facing the Kinect© and make a move to the left or right, quickly.
- Zoom** This gesture is to put the palm facing the Kinect© and make a move down quickly.
- Close Window** This gesture is a change of direction of the palm, on the X axis, four times.



Fig. 2. Cinveswall and Kinect©

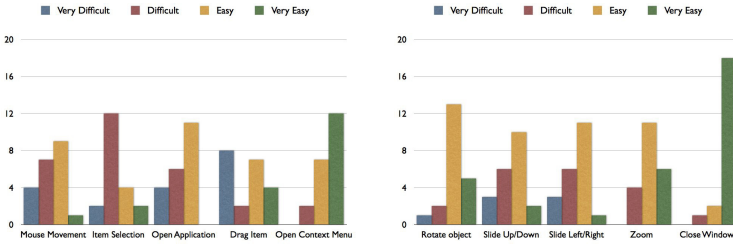


Fig. 3. Usability charts

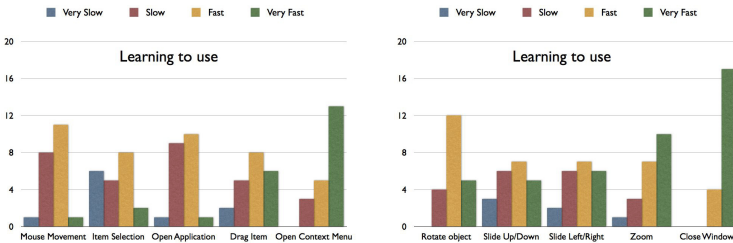


Fig. 4. Charts of user training

3.2 Tools and Equipment

The gestures considered for the second test (use of applications and desktop) are:

Select character in the virtual keyboard It involves moving the mouse pointer to the button of the character and click.

Write text This gesture includes selecting a text field, and then select the character set of the text.

Write number This gesture includes selecting a text field, and then select the numeric character set of the number.

Delete text Position the pointer to the left of text character to delete and press the delete key.

CinvesWall. The CinvesWall is an array of 12 Apple Cinema Display 24-inch, in a 3x4 configuration, controlled by a cluster of 12 Mac mini and one MacPro server as the master of visualization cluster, figure 2. The network visualization cluster interconnect is Gigabit Ethernet. The total resolution of the display device is 27 megapixels (7680x3600). The nodes in the cluster work with OSX version 10.6.5 (Snow Leopard). The administration of the nodes is done with the application Apple Remote Desktop.

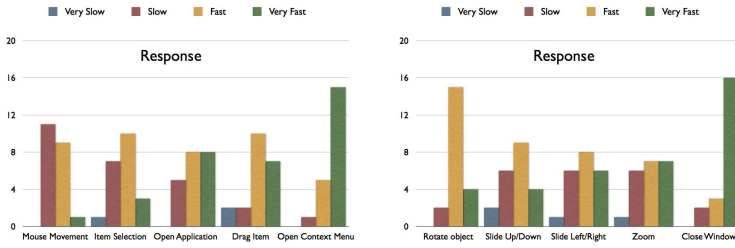


Fig. 5. Performance charts

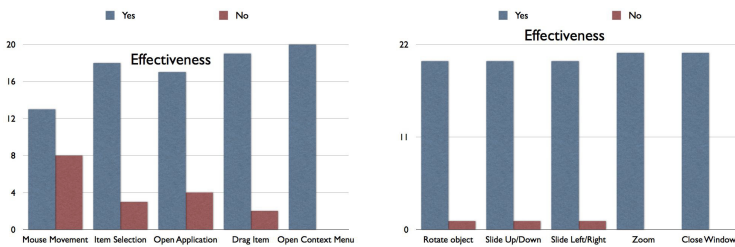


Fig. 6. Effectiveness charts

Kinect©Setup. The Kinect© control the user gestures. The Kinect is connected to the master node, but is placed in the middle of the video wall to give the illusion that the user interacts directly with the high-resolution display. In order for users to interact with the video wall, standing five feet away from the Kinect©, as show in Figure 2.

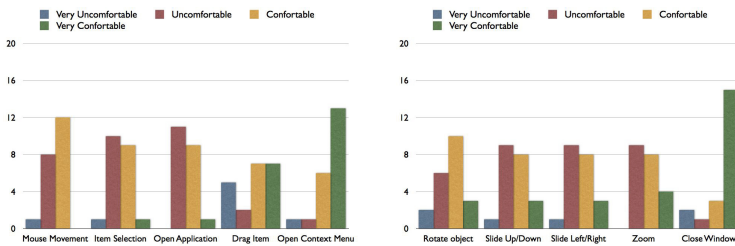


Fig. 7. Charts of user comfort

Applications Settings. Tests of actions and tasks, with the Kinect[®], were performed using distributed applications running on tiled display and applications running on the distributed desktop. Fortunately we were able to use an application that runs in both formats: ParaView. We also use PreView application running on the OSX desktop and equivalent distributed image display application. Several tasks are able to test in the context of execution of an application. The rest of the tasks and actions were used to manage the distributed desktop.

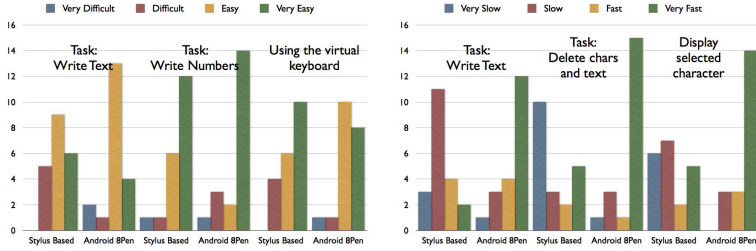


Fig. 8. Usability and performance charts of use the virtual keyboard

4 Results

Since we want to measure the use of Kinect[®] as interaction device with video wall, we intend measure usability, effectiveness, performance and comfort. Started showing the results of applying the tests to the tasks and actions. And then see the results of tests using virtual keyboards.

4.1 Results for Test on Actions and Tasks

The chart in the Figures 3 and 4 displays the usability and user training results. This training is related to learn and associate a gesture to an action or task.

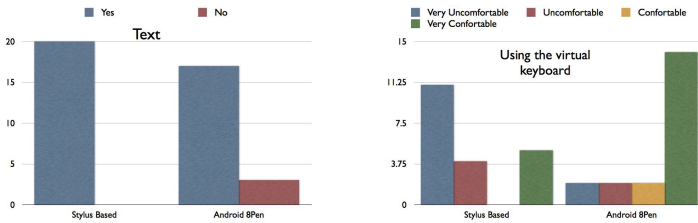


Fig. 9. Effectiveness chart and chart of user comfort using the virtual keyboard

We can measure the performance of the Kinect© using the time of respond to action or task. The Figure 5 show the user's point of view about the response. The effectiveness of actions and task performed by the users throught Kinect© is showed in Figure 6. Figure 5 show the user's point of view about the response. The ease of use of gesture for Kinect, shown in Figure 7.

4.2 Test Results of the Use of Virtual Keyboards

The Figure 8 show the charts of usability and performance of virtual keyboards: Stylus based and Android 8Pen. And figure 9 show the effectiveness chart and chart of user comfort using the virtual keyboard.

5 Discussion

The configuration of tiled displays based on visualization server use the homogeneous visualizations capabilities in every output of the graphics cards. This configuration take advantage that the response to user actions are very quick. In this work we are using a distributed control to the tiled display. When the main task is to visualize these devices with high-resolution images, the user might be tolerant synchronization waits and displays. However, when the main task is the interaction, the user waits for answers in a very short time. The Cinveswall use a Gigabit Ethernet switch that does not allow multicast operation. This defect causes users do not have such good results in the tasks associated with drag the mouse pointer and selection.

6 Conclusions

The interaction use of the tiled display need that the time of the graphics response is in the range accepted for the user. In the case of having a visualization cluster, you must ensure rapid communication between different nodes, so that the user has the idea that the response is adequate in the tiled display.

There is a relationship between the area of deployment of the video wall and the step size of the cursor on the remote desktop. The relationship is that the higher the display area, the step size should be smaller.

In testing, you can verify that the user's interaction with applications running in a distributed, as ParaView and image viewer, is similar to the interaction of the applications that run on the desktop distributed as ParaView and Preview. For the type of results obtained in this work. Best tasks associated gestures using the Kinect, are not related to the use of the mouse pointer.

Finally, the use of virtual keyboards for handling text fields seems appropriate. Surprisingly, 8pen Android appears to offer advantages over the Stylus based.

Acknowledgements. The authors wish to thank Arellano Cenizeros Israel, González Hernández Carlos Omar, Martínez Martínez Jesús Oswaldo and Julia Leticia Sánchez Sánchez for their valuable assistance in the implementation of the tests were performed.

References

- [TS] Tao, N., Schmidt, G.S., Staadt, O.G., Livingston, M.A., Ball, R., May, R.: A Survey of Large High-Resolution Display Technologies, Techniques, and Applications. In: Proceedings of the IEEE Conference on Virtual Reality, vol. 236, pp. 223–236. IEEE Computer Society, Washington, DC (2006)
- [SC] Smarr, L.L., Chien, A.A., DeFanti, T., Leigh, J., Papadopoulos, P.M.: The OptIPuter. *Communications of the ACM* 14(11), 58–66 (2003)
- [MC] Viveros, A.M., Vergara, S.V.C.: The CinvesWall in More than Research. In: Proceedings of the 2nd International Supercomputing Conference in México, ISUM 2011, vol. 2, pp. 177–183. Universidad de Guadalajara, México (2011)
- [NH] Nirnimesh, H.P., Narayanan, P.J.: Garuda: A Scalable Tiled Display Wall Using Commodity PCs Visualization and Computer Graphics. *IEEE Transactions on* 13(5), 864–877 (2007)
- [NV] Naveen, K., Venkatram, V., Vaidya, C., Nicholas, S., Allan, S., Charles, Z., Gideon, G., Jason, L., Andrew, J.: SAGE: the Scalable Adaptive Graphics Environment. In: WACE (2004)
- [CC] Chen, Y., Chen, H., Clark, D.W., Liu, Z., Wallace, G., Li, K.: Software Environments for Cluster-based Display Systems. In: First IEEE/ACM International Symposium on Cluster Computing and the Grid (2001)
- [ET] Ebert, A., Thelen, S., Olech, P.-S., Meyer, J., Hagen, H.: Tiled++: An Enhanced Tiled Hi-Res Display Wall. *IEEE Transactions On Visualization and Computer Graphics* 16(1) (2010)
- [YN] Yost, B., North, C.: The perceptual scalability of visualization. *IEEE Transactions On Visualization And Computer Graphics* 12(5) (September/October 2006)
- [BN1] Ball, R., North, C.: Analysis of User Behavior on High-Resolution Tiled Displays. In: Costabile, M.F., Paternó, F. (eds.) *INTERACT 2005*. LNCS, vol. 3585, pp. 350–363. Springer, Heidelberg (2005)
- [BN2] Ball, R., North, C.: Effects of tiled high-resolution display on basic visualization and navigation tasks. In: *CHI 2005 Extended Abstracts on Human Factors in Computing Systems (CHI EA 2005)*, pp. 1196–1199. ACM, New York (2005)
- [AT] Ahlborn, B.A., Thompson, D., Kreylos, O., Hamann, B., Staadt, O.G.: A practical system for laser pointer interaction on large displays. In: Proceedings of the ACM Symposium on Virtual Reality Software and Technology (VRST 2005), pp. 106–109. ACM, New York (2005)
- [AG] Israel, A.C., Omar, G.H.C., Oswaldo, M.M.J.: Control de un VideoWall via Kinect®. In: México, D.F. (ed.) *Trabajo Terminal, Escuela Superior de Cómputo, ESCOM-IPN, Junio 2012* (2012)
- [RCMM] Ramírez, L., Chapa, S., Meneses, A.: DVO: Model for Make a Handler for a Tiled Display. *Lecture Notes in Engineering and Computer Science*, vol. 2198(1), pp. 995–1001 (2012)
- [MH] Myers, B., Hudson, S.E., Pausch, R.: Past, present, and future of user interface software tools. *ACM Trans. Comput.-Hum. Interact.* 7(1), 3–28 (2000)
- [PR] Puder, A., Romer, K., Pilhofer, F.: *Distributed Systems Architecture, A Middleware Approach*. Elsevier Inc. (2006)
- [HH] Ng, Y., Humphreys, G., Houston, M.: Chromium: A stream processing framework for interactive rendering on clusters. *ACM TOG* 21(3) (2002)

- [GdS] Germans, D., van der Schaaf, T., Renambot, L., et al.: Retained mode parallel rendering for scalable tiled displays. In: IPT (2002)
- [R] Patricia, L., Rivera, R.: “Minería de datos visual sobre una pared de video”, Master Degree Thesis, Departamento de Computación, CINVESTAV-IPN (2008)
- [BL] Granados, G.A.B., García, A.L.L.: Manejo de video distribuido sobre un Video Wall, Trabajo Terminal, Escuela Superior de Cómputo del IPN, Junio de (2009)
- [TGSP] Tan, D.S., Gergle, D., Scupelli, P.G., Pausch, R.: Physically large displays improve path integration in 3D virtual navigation tasks. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2004 (2004)
- [BG] Birnholtz, J.P., Grossman, T., Mak, C., Balakrishnan, R.: An exploratory study of input configuration and group process in a negotiation task using a large display. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2007). ACM, New York (2007)
- [NJ] Nam, S., Jeong, B., Renambot, L., Johnson, A., Gaither, K., Leigh, J.: Remote visualization of large scale data for ultra-high resolution display environments. In: Ma, K.-L., Papka, M.E. (eds.) Proceedings of the 2009 Workshop on Ultra-scale Visualization (UltraVis 2009), pp. 42–44. ACM, New York (2009)
- [KB] Boulos, M.N.K., Blanchard, B.J., Walker, C., Montero, J., Tripathy, A., Gutierrez-Osuna, R.: Web GIS in practice X: a Microsoft Kinect natural user interface for Google Earth navigation. *International Journal of Health Geographics* 2011 10(45) (2011)
- [MRCC] Microsoft Research Cambridge & Xbox Incubation, Real-Time Human Pose Recognition in Parts from Single Depth Images

Study on Cursor Shape Suitable for Eye-gaze Input System

Atsuo Murata, Raku Uetsugi, and Takehito Hayami

Graduate School of Natural Science and Technology, Okayama University, Okayama, Japan
{murata,uetsugi}@iims.sys.okayama-u.ac.jp)

Abstract. The aim of this study was to identify the cursor shape suitable for eye-gaze interfaces. The conventional arrow shape was, irrespective of the number of targets in the display, not suitable for an eye-gaze input system from the perspective of task completion time, number of errors, and subjective rating on usability. It is recommended that the cursor shape of an eye-gaze input system should be cross or ellipse. When the distance between targets is wider, the ellipse type is proper.

Keywords: cursor shape, speed, accuracy, eye-gaze input system.

1 Introduction

The technology for measuring a user's visual line of gaze in real time has been advancing. Appropriate human-computer interaction techniques that incorporate eye movements into a human-computer dialogue has been developed [1-9]. These studies have found the advantage of eye-gaze input system. However, few studies except Murata [8] have examined the effectiveness of such systems with older adults. Murata [8] discussed the usability of an eye-gaze input system to aid interactions with computers for older adults. Systematically manipulating experimental conditions such as the movement distance, target size, and direction of movement, an eye-gaze input system was found to lead to faster pointing time as compared with mouse input especially for older adults.

As eye-gaze input interfaces enable us to interact with PC by making use of eye movements, it is expected that even disables persons with deficiency on the upper limb can easily use it. A lot of studies are reported on eye-gaze input interfaces as an alternative to a mouse. However, there are still a few problems we must overcome so that such an input system can be put into practical use.

The shape of mouse cursor suitable for general human-computer interactions (HCI) except for eye-gaze interfaces is discussed, for example, by Pastel [10], Lecquier [11], and Phillips [12]. Like general HCI, we should use a proper cursor shape which enhances the usability of eye-gaze input system. As the eye-gaze input system differs from the mouse input in input mechanism, and has a lower resolution as compared with the mouse input, it is natural and reasonable to predict that the cursor shape

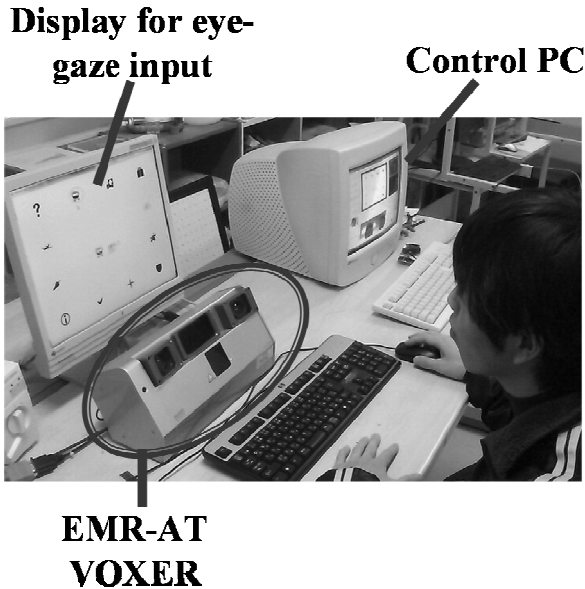


Fig. 1. Eye camera used in the experiment

proper for the mouse input does not necessarily lead to the high usability of eye-gaze interfaces. Although a conventional arrow-type cursor is used even in eye-gaze input interface, there seems to be no definite reason to use such a conventional cursor shape in eye-gaze input interfaces.

Until now, it has not been explored what type of cursor shape is suitable for eye-gaze input interfaces. The aim of this study was to identify the cursor shape suitable for eye-gaze interfaces.

2 Method

2.1 Participant

Ten healthy young adults aged from 21 to 24 years old took part in the experiment. All participants had an experience of personal computer with an average of 5.5 years (6-7 years). The visual acuity of the participants in both young and older groups was matched and more than 20/20. They had no orthopedic or neurological diseases.

2.2 Apparatus

Using EMR-AT VOXER (Nac Image Technology) (See Fig.1), an eye-gaze input interface was developed. Visual C# (Microsoft) was used as a programming language. This apparatus enables us to determine eye movements and fixation by measuring the reflection of low-level infrared light (800 nm), and also admits the head movements

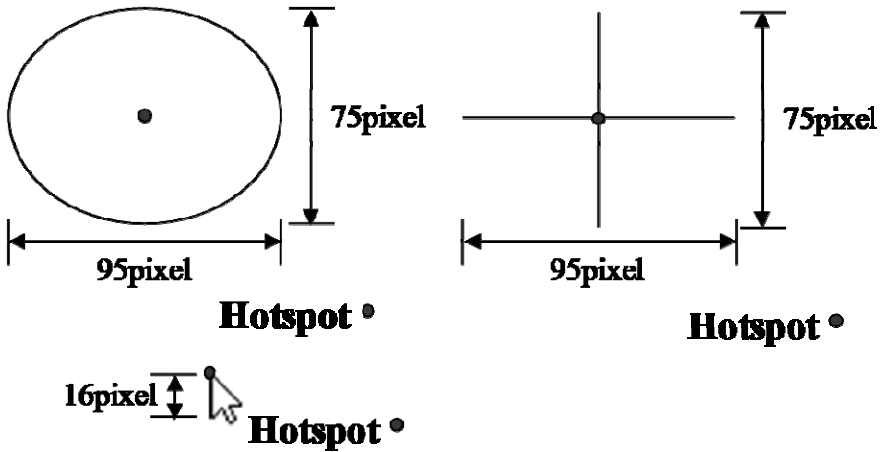


Fig. 2. Explanation three types of cursor shape

within a predetermined range. The eye-tracker was connected with a personal computer (HP, DX5150MT) with a 15-inch (303mm x 231mm) CRT. The resolution was 1024 x 768pixel. Another personal computer was also connected to the eye-tracker via a RS232C port to develop an eye-gaze input system. The line of gaze, via a RS232C port, is output to this computer with a sampling frequency of 60Hz. The illumination on the keyboard of a personal was about 200lx, and the mean brightness of 5 points (four edges and a center) on CRT was about 100cd/m². The viewing distance was about 70 cm.

2.3 Cursor Shape

In this study, the cursor shape was focused on, and selected as an within-subject experimental factor. The number of icons to be searched for was also a within-subject experimental variable. The usability was compared among the ellipse-type cursor, the cross-type cursor, and the conventional arrow-type cursor in order to clarify the cursor shape suitable for eye-gaze input interfaces. The types of cursor are explained in Fig.2. The number of icons displayed on CRT (12, 32, and 60 targets) and the corresponding interval between icons to be pointed to was also controlled as an experimental factor. The evaluation measures were: task completion time, number of errors, and subjective rating on usability for each shape of cursor. The three types of displays used in the experiment are depicted in Fig.3(a)-(c).

2.4 Design and Procedure

The participants performed a total of nine conditions (three shapes (arrow, cross, and ellipse) by three numbers of icons (12, 32, and 60)). The order of performance of nine experimental conditions was randomized across the participants. After the cali bration

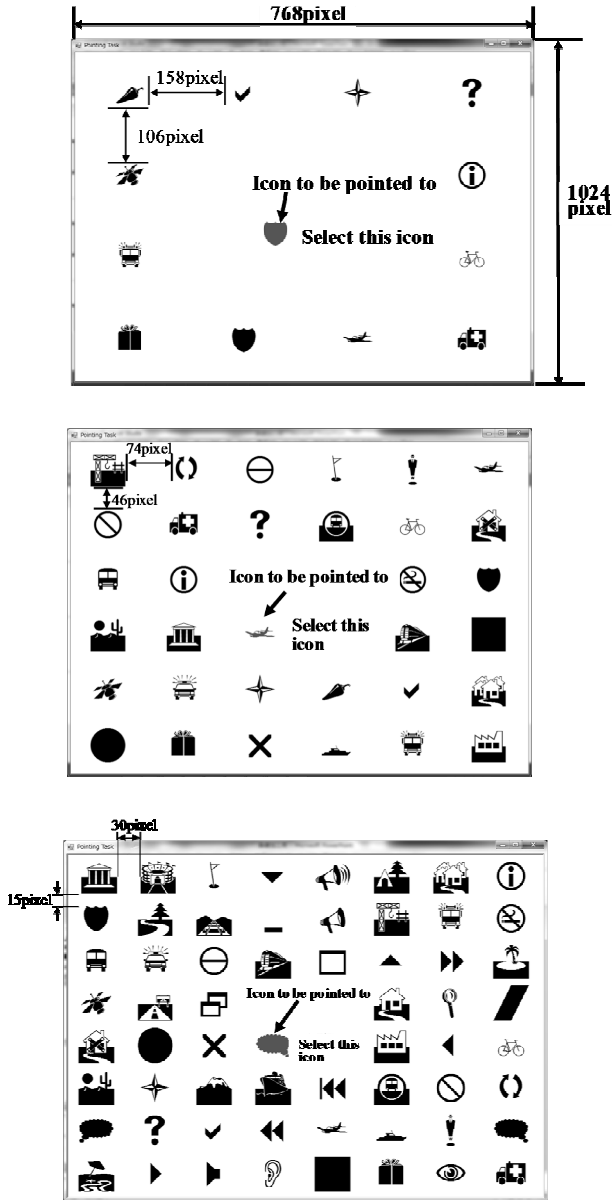


Fig. 3. Display used in the experiment. (a) 12 icons, (b) 32 icons, (c) 60 icons.

of eye camera and the practice session, the participants entered the experimental session. First, the participants were ordered to fixate the center of the display. After the fixation to the center area, the icon to be pointed to is presented on the display in Fig.1. The participants move their fixation from the central area to the specified icon, and fixate there for the predetermined duration. This corresponds to one pointing trial.

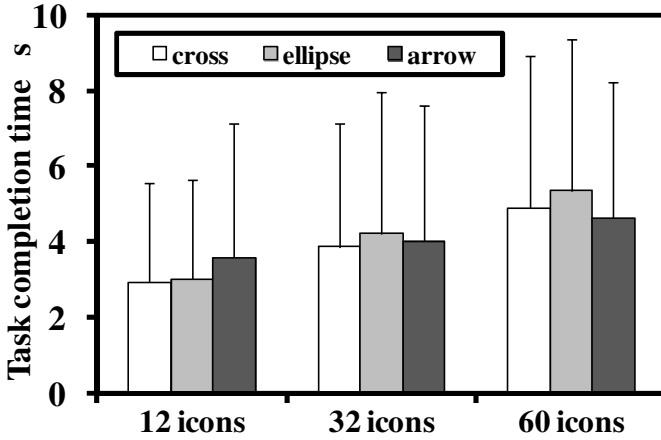


Fig. 4. Task completion time as a function of cursor shape and number of icons

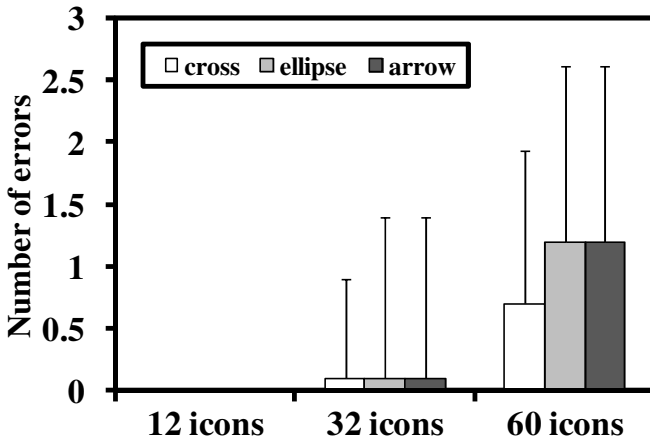


Fig. 5. Number of errors as a function of cursor shape and number of icons

For one experimental session, the participants were required to carry out 10 pointing trials.

3 Results

In Fig.4, the mean pointing time (task completion time) is plotted as a function of cursor shape and number of icons. The pointing time of the ellipse-type and the cross-type cursors was shorter than that of the conventional arrow-type cursor when targets were fewer and the interval between targets was wider. In Fig.5, the number of errors is shown as a function of cursor shape and number of icons. As a whole, the cross - type cursor suffered from few errors. In Fig.6, the task completion time is compared

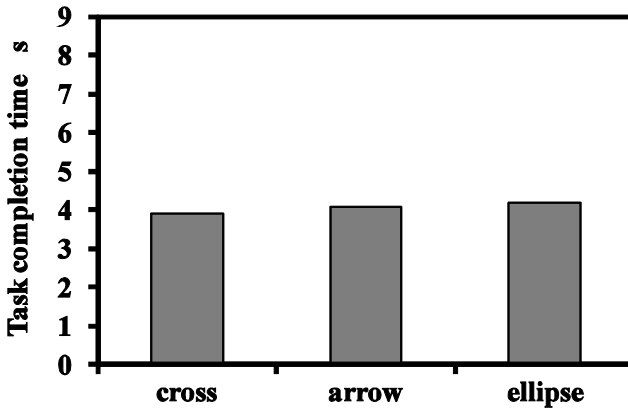


Fig. 6. Task completion time as a function of cursor shape

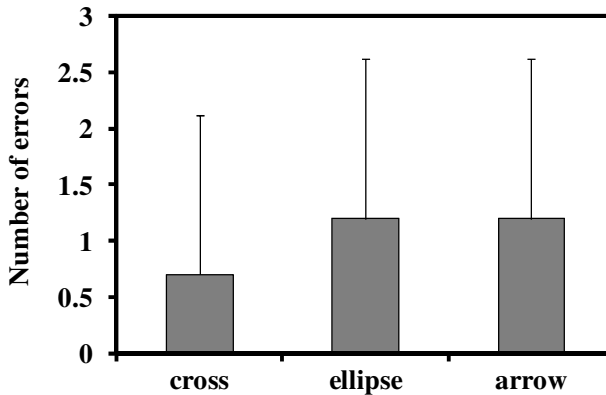


Fig. 7. Number of errors as a function of cursor shape

among three cursor shapes. In Fig.7, the number of errors is compared among three cursor shapes. In Fig.8, the subjective rating on usability is compared among three types of cursor shapes. Fig.9 compares the subjective rating on fatigue among three cursor types. The subjective rating on usability for the ellipse-type cursor was the highest. With the increase of the number of targets, however, the pointing time of the ellipse-type and the cross-type cursors tended to be prolonged.

A one-way (cursor shape) ANOVA (Analysis of Variance) carried out on the pointing time revealed no significant main effect of cursor shape. A similar one-way ANOVA conducted on the number of error also revealed no significant main effect of cursor shape. A two-way (cursor shape and number of targets) ANOVA conducted on the pointing time revealed only a significant main effect of number of targets ($F(2,27)=13.739, p<0.01$). A similar two-way ANOVA carried out on the number of error detected only a significant main effect of number of targets ($F(2,27)=11.870, p<0.01$).

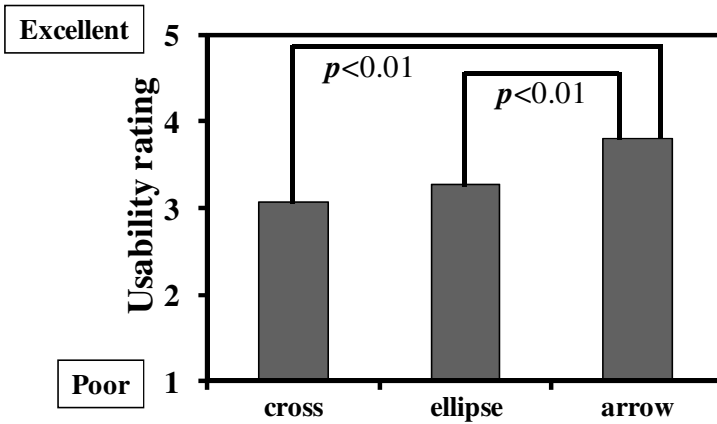


Fig. 8. Subjective rating on usability as a function of cursor shape

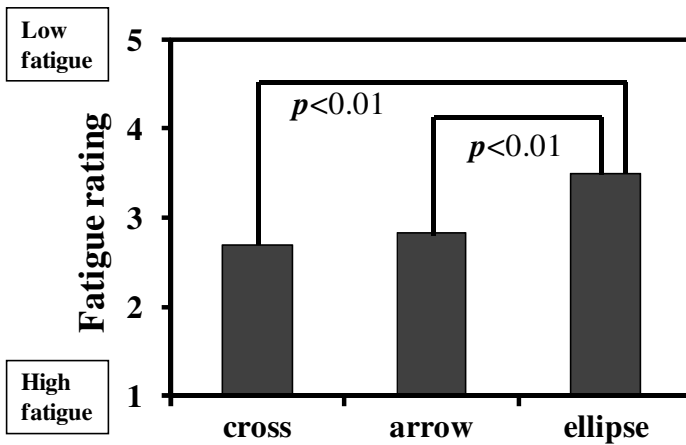


Fig. 9. Subjective rating on fatigue as a function of cursor shape

A Bonferroni/Dunn non-parametric test was carried out on the rating score on usability. As a result, the following significant differences ($p < 0.01$) were detected: (ellipse, arrow) and (ellipse, cross). A similar non-parametric test carried out on the fatigue rating also revealed the following significant differences: (ellipse, arrow) and (ellipse, cross).

4 Discussion

From Figs.8 and 9, the findings can be summarized as follows. The ellipse cursor was found to be highly evaluated from the aspects of both ease of operation and fatigue. As the hotspot of ellipse cursor is invisible, one can operate without consciously recognizing a cursor. This must lead to lower fatigue rating. The cross cursor got the

worst evaluation on ease of operation and fatigue. The larger cursor size of cross type makes one feel the fluctuation of cursor smaller. As the hotspot is definitely visible in the cross cursor, this must lead to the lower rating score on both ease of operation and fatigue. The arrow cursor did not get higher evaluation on ease of operation and fatigue. This means that the experience of using a mouse cursor (arrow cursor) does not affect the usability of eye-gaze input system. In other words, the experience of arrow cursor when using a mouse does not necessarily influence positively the usability of an eye-gaze input system.

In conclusion, it was confirmed that the conventional arrow shape was not suitable for an eye-gaze input system from the viewpoints of pointing time (task completion time), number of errors, and subjective rating on usability (see Figs.4-7). It is recommended that the cursor shape of an eye-gaze input system should be cross or ellipse. In more detail, the cross cursor is appropriate in the aspect of both operation speed and accuracy. Taking the psychological rating into account, the ellipse cursor is strongly recommended so long as the distance between targets is wide enough like the condition of 12 icons in this study.

References

1. Jacob, R.J.K.: What you look at is what you get: Eyemovement- based interaction technique. In: Proceedings of ACM CHI 1990, pp. 11–18 (1990)
2. Jacob, R.J.K.: The use of eye movements in human-computer interaction techniques: What you look at is what you get. *ACM Transactions on Information Systems* 9, 152–169 (1991)
3. Jacob, R.J.K.: Eye-movement-based human-computer interaction techniques: Towards non-command interfaces. In: Harston, H.R., Hix, D. (eds.) *Advances in Human-Computer Interaction*, vol. 4, pp. 151–190. Ablex, Norwood (1993)
4. Jacob, R.J.K.: What you look at is what you get: Using eye movements as computer input. In: *Proceedings of Virtual Reality Systems 1993*, pp. 164–166 (1993)
5. Jacob, R.J.K.: Eye tracking in advanced interface design. In: Baefield, W., Furness, T. (eds.) *Advanced Interface Design and Virtual Environments*, pp. 212–231. Oxford University Press, Oxford (1994)
6. Jacob, R.J.K., Sibert, L.E., Mcfarlanes, D.C., Mullen, M.P.: Integrality and reparability of input devices. *ACM Transactions on Computer-Human Interaction*, 2–26 (1994)
7. Sibert, L.E., Jacob, R.J.K.: Evaluation of eye gaze interaction. In: *Proceedings of CHI 2000*, pp. 281–288 (2000)
8. Murata, A.: Eye-gaze input versus mouse: cursor control as a function of age. *International Journal of Human-Computer Interaction* 21, 1–14 (2006)
9. Murata, A., Moriwaka, M.: Effectiveness of the menu selection method for eye-gaze input system -Comparison between young and older adults. In: *Proceedings of International Workshop on Computational Intelligence and Applications, IWCA 2009*, pp. 306–311 (2009)
10. Pastel, R.: Positioning graphical objects on computer screens: A three-phase model. *Human Factors* 53(1), 22–37 (2011)
11. Lecquier, A.: A study of the modification of the speed and size of the cursor for simulating pseudo-haptic bumps and holes. *ACM Transactions on Applied Perception* 5(3) (2008)
12. Phillips, G.: Conflicting directional and locational cues afforded by arrowhead cursors in graphical user interfaces. *Journal of Experimental Psychology: Applied* 9(2), 75–87 (2003)

Study on Character Input Methods Using Eye-gaze Input Interface

Atsuo Murata, Kazuya Hayashi, Makoto Moriwaka, and Takehito Hayami

Graduate School of Natural Science and Technology,
Okayama University, Okayama, Japan
{murata,moriwaka}@iims.sys.okayama-u.ac.jp

Abstract. Four character input methods for eye-gaze input interface were compared from the viewpoints of input speed, input accuracy, and subjective rating on ease of input and fatigue. Four input methods included (1) I-QGSM (vertical), (2) I-QGSM (circle), (3) eye-fixation method, and (4) screen button. While the eye-fixation method (3) led to faster input, the I-QGSM (vertical) led to fewer errors. In conclusion, it is difficult to develop character input method that satisfies both speed and accuracy.

Keywords: Character input, eye-gaze input system, I-QGSM, eye-fixation, speed, accuracy.

1 Introduction

Older people present an increasingly large portion of the population and are likely to be active users of IT. Issues surrounding IT and aging are, therefore, of much interest to not only researchers but also practitioner within the domain of human-computer interaction (HCI). Therefore, the development of an input device that is friendly to older adults and leads to higher performance is essential. There are many reports suggesting that older adults exhibit deficits in various cognitive motor tasks [1,2]. Spatial abilities, that is, the capacity to acquire, manipulate, and use information on Web pages, have been shown to decline with age[3], and this might account for the difficulties of older adults when navigating Web pages. Kelly and Charness [4] showed that spatial abilities may be important for mediating the effects of age on computing skills. Processing speed refers to the ability to acquire, interpret, and respond to information quickly and accurately. Salthouse [5] pointed out that reductions in processing speed are a common explanation for many age-related deficits in task performance. Therefore, it is expected that decreasing motor function in older adults hinders the successful use of input devices such as a mouse and generally leads to a relatively longer pointing time and lower pointing accuracy in comparison with young counterparts.

The technology for measuring a user's visual line of gaze in real time has been advancing. Appropriate human-computer interaction techniques that incorporate eye movements into a human-computer dialogue has been developed [6-15]. These studies

have found the advantage of eye-gaze input system. However, few studies except Murata [13] have examined the effectiveness of such systems with older adults. Murata [13] discussed the usability of an eye-gaze input system to aid interactions with computers for older adults. Systematically manipulating experimental conditions such as the movement distance, target size, and direction of movement, an eye-gaze input system was found to lead to faster pointing time as compared with mouse input especially for older adults.

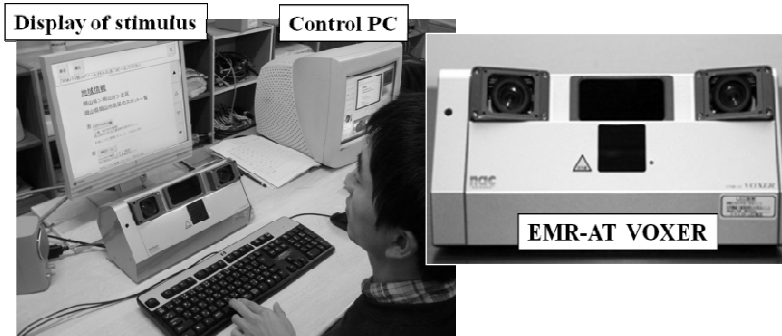


Fig. 1. Eye camera used in the experiment

However, the studies above [6-15] cannot be applied to the real-world computer systems such as Internet Explorer. The character input method suitable for an eye-gaze input interface is essential to make it possible for such an interface to be used practically on Web applications such as Internet Explorer. However, there are few studies that examined an effective input method using an eye-gaze input interface. The input methods proposed until now suffer from the problem of error input of characters. Frequent input errors keep users from using such an interface, and brake the widespread of it. The aim of this study was to realize (program) four kinds of input methods, examine the usability of these systems, and propose an input method suitable for an eye-gaze input interface. Four methods includes: (1)I-QGSM (vertical), (2)I-QGSM (circle), (3)Eye fixation, and (4)Screen button.

2 Method

2.1 Participants

Five healthy young adults aged from 22 to 24 years took part in the experiment. All participants had an experience of personal computer with an average of 5.5 years (6-7 years). The visual acuity of the participants in both young and older groups was matched and more than 20/20. They had no orthopedic or neurological diseases.

2.2 Apparatus

Using an eye camera shown in Fig.1, the eye-gaze input system was developed. This apparatus enables us to determine eye movements and fixation by measuring the

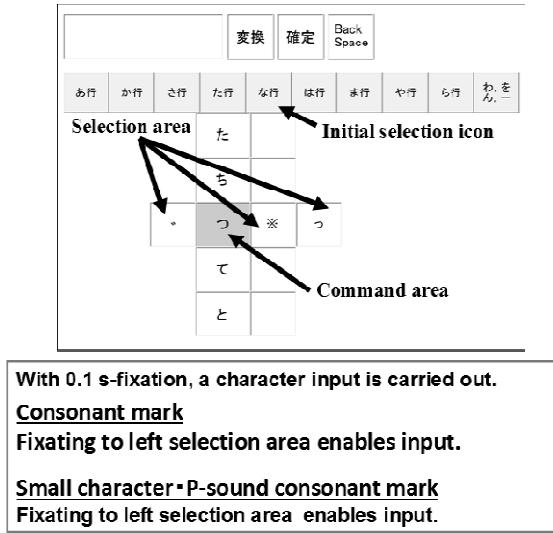


Fig. 2. I-QGSM(Vertical) input method

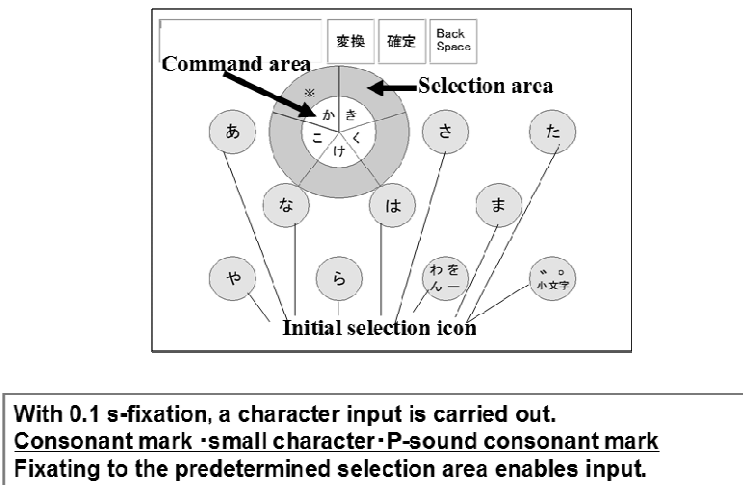


Fig. 3. QGSM(Circle) input method

reflection of low-level infrared light (800 nm), and also admits the head movements within a predetermined range. The eye-tracker was connected with a personal computer (HP, DX5150MT) with a 15-inch (303mm x 231mm) CRT. The resolution

was 1024 x 768pixel. Another personal computer was also connected to the eye-tracker via a RS232C port to develop an eye-gaze input system. The line of gaze, via a RS232C port, is output to this computer with a sampling frequency of 60Hz. The illumination on the keyboard of a personal was about 200lx, and the mean brightness of 5 points (four edges and a center) on CRT was about 100cd/m². The viewing distance was about 70 cm.

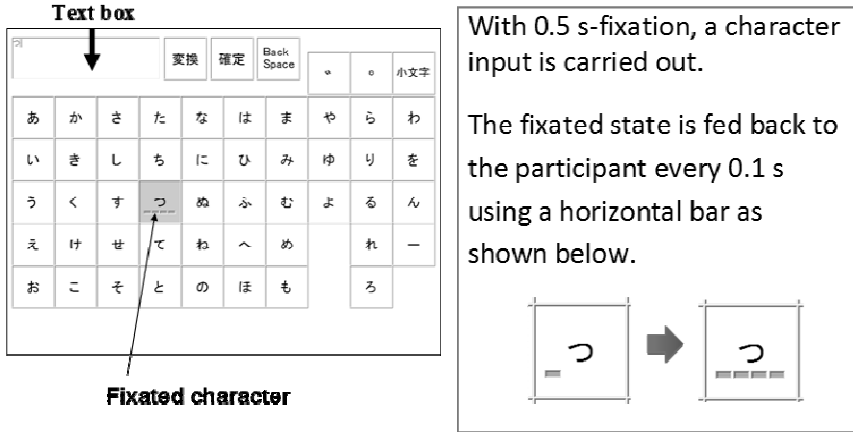


Fig. 4. Eye-fixation input method

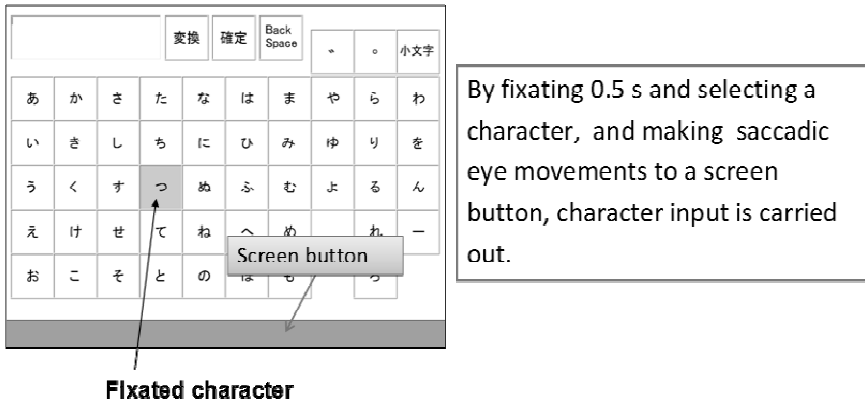


Fig. 5. Screen button method

2.3 Character Input Method

Two kinds of character input method for the eye-gaze input interface using I-QGSM (Improved-Quick Glance Selection Method) reported in Murata et al. [14] were proposed. The effectiveness was experimentally compared among the proposed method (vertical display) (method (1)), the proposed method (circle display) (method (2)), the

conventional eye-fixation method (method (3)), and the conventional screen-button pressing method (method (4)). The displays of methods (1)-(4) are depicted in Fig.2-Fig.5, respectively.

The four methods are briefly explained as follows:

(1) I-QGSM (vertical) (Fig.2): After the fixation to the initial selection icon for more than 0.2 s, the command area and the selection area appears as in Fig.1. Eye movements to a necessary area for an input enable us to input a character.

(2) I-QGSM (circle) (Fig.3): In a similar way to the method (1), a character is entered.

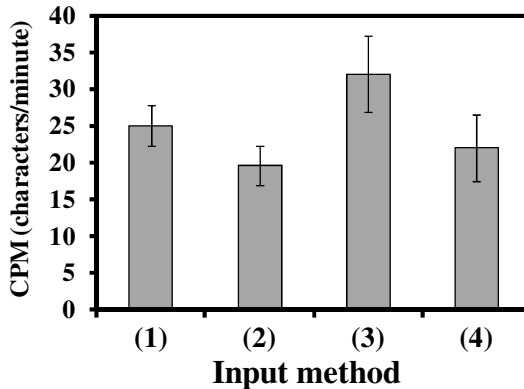


Fig. 6. Characters/min (CPM) as a function of input method

(3) Eye fixation (Fig.4): The fixation to a character which the participant needs to enter completes a character input. This procedure is repeated.

(4) Screen button: (Fig.5): After focusing a character by fixating to it for 0.1 s and moving an eye-gaze to the screen button in Fig.5, the input is completed by.

2.4 Design and Procedure

The participant was required to input 10 words of 5-8 characters for each input method. This was repeated two times. Only the data of second session was used for the analysis. A total number of characters were 60. The participant was permitted to have a short break between sessions. After one word had been finished inputting, the next word appeared on the display. After the input of 120 characters had been finished (two sessions had been finished), the participant was required to evaluate the ease of input and the fatigue with a five-point scale for each input method. The order of performance of four input methods was randomized across the participants.

2.5 Evaluation Measures

The valuation measures were entered characters per one minute (CPM) and the number of errors. The task was to enter a total of 60 words of 5-8 characters. The psychological ratings on ease of input and fatigue induced during the experimental task were also used as evaluation measures.

3 Results

3.1 Characters Per Minute (CPM)

In Fig.6, entered characters per one minute (CPM) are compared among four input methods. As a result of a one-way ANOVA (Analysis of Variance) conducted on the CPM, a significant main effect of input method ($F(3,16)=19.6, p<0.01$) was detected.

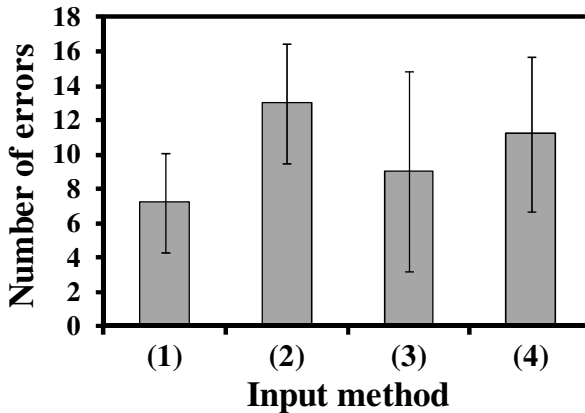


Fig. 7. Number of errors as a function of input method

3.2 Number of Errors

In Fig.7, the number of entry errors is compared among four input methods. A similar one-way ANOVA conducted on the number of entry errors did not detect a significant main effect.

3.3 Rating on Ease of Input and Fatigue

In Fig.8, the rating score on ease of input is plotted as a function of input method. Fig.9 compares the rating score on fatigue among four input methods (1)-(4). Kruskal-Wallis non-parametric test conducted on the rating score of ease of input revealed no significant main effect of input method. A similar non-parametric test carried out on the fatigue rating also revealed no significant main effect of input method.

4 Discussion

4.1 Pointing Speed

The conventional eye-fixation method led to the quickest input. The reason can be inferred as follows. The eye-fixation method needed the fewest eye movements of the four methods. Although the entry speed was the fastest, the entry errors occurred

more frequently than I-QGSM (vertical) due to a shorter fixation time. Together with such tendencies, the rating score on ease of input for (3) eye-fixation method tended to be higher. The rating score on fatigue for (3) eye-fixation method was evaluated highly. The rating score for (1) I-QGSM(vertical) was moderate, although this led to the fewest errors of all of four input methods. On the basis of this result, it might be possible that the participant psychologically evaluate highly for the input method that enables him to input characters fast.

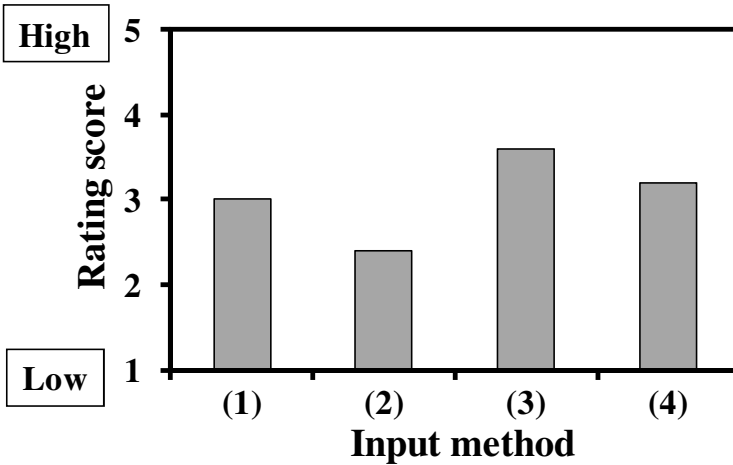


Fig. 8. Rating score on ease of input as a function of input method

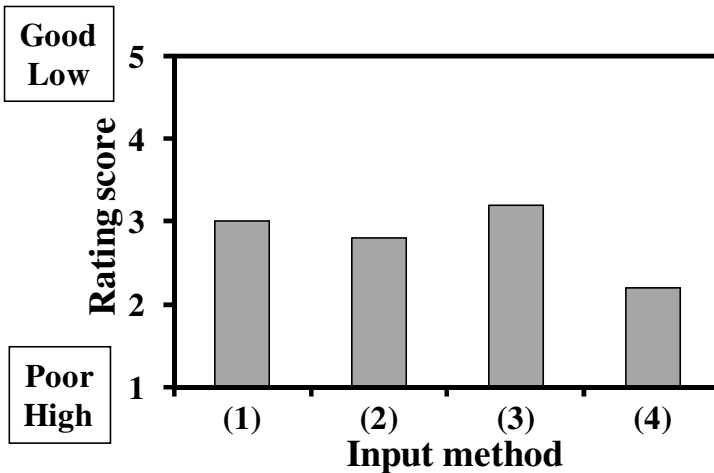


Fig. 9. Rating score on fatigue as a function of input method

4.2 Pointing Accuracy

As for (1) I-QGSM (vertical), the participant must gaze at the initial selection icon. Then, the participant sequentially gazes at the vertical column. Therefore, it takes time for the participant to input a character located at the bottom. Therefore, the input speed in this method must be slower as compared with that of (3) eye-gaze input. In this method (1), input is impossible if the participant does not gaze at the selection area. This must lead to fewer input errors. For the case of input of consonant mark, the input method was simplified. When entering a consonant mark, the following errors occurred: errors in movement direction and over-movement (See Fig.10). Such errors must be reduced in future development.

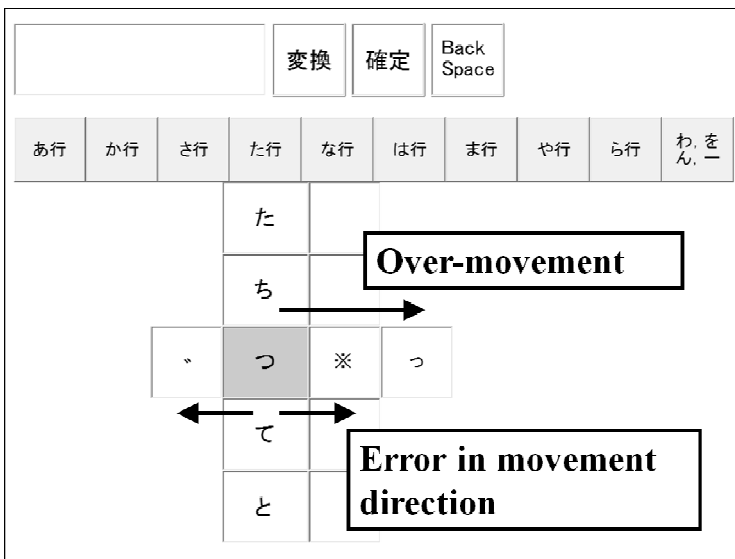


Fig. 10. Error frequently occurred during the input of consonant mark in (1) I-QGSM

It is difficult to grasp the location of character when using (2)I-QGSM (circle). The oblique movement is difficult in this method. The different character group was frequently selected, which must lead to more input errors and slower input speed.

In (3) eye-fixation method, the distance to the next characters is minimal. Consequently, the eye movements necessary for the input were least. The simplest character input was possible. Low fixation duration is also possible. Therefore, these must lead to the fastest input speed. Due to the shorter fixation duration, input errors were observed to some extent. This issue must be further explored in future work.

As for (4) screen button, the following problems were identified. As the participant must gaze at the screen button every character input, the eye movement is frequently done. Due to un-familiarized saccadic eye movement, input errors frequently occurred (The focus is moved to other character until the movement to screen button is completed. These problems must be improved in future development.

4.3 Psychological Rating on Usability and Fatigue

As shown on the basis of performance data such as entered characters per minute (CPM) and number of entry errors, the input method (3) led to better performance. In accordance with this, the subjective rating on ease of input and induced fatigue during the task for the input method (3) tended to be higher. As for the eye fixation (input method (3)), it has also shown that the click operation using an eye-gaze input system is more speedy and accurate when the eye fixation was utilized (Murata et al.[14]). The eye-gaze input system that made use of the eye fixation also obtained higher psychological evaluation on usability.

On the basis of the three viewpoints, it seems that the input method (3) based on the eye fixation is promising for the input means.

4.4 Implication for Designing HCI of Eye-gaze Input System

For eye-gaze input, the following opinions were reported.

- As fixation leads to the input of character, the participant must always carry out eye movements.
- Such a situation is felt to be time pressured.

As for the pointing accuracy, the proposed method (vertical display) was found to be free from errors. It seems difficult to develop an input method which satisfies both speed and accuracy. In conclusion, we should make the proper use of the input methods according to the criterion (speed or accuracy) on which an emphasis is placed. While (3)Eye-fixation enables us to input characters fast, (1)I-QGSM (vertical) ensure input with fewer errors. Future work should be done to reduction input errors. Moreover, the habituation process of each input system should be explored.

References

1. Goggin, N.L., Stelmach, G.E., Amrhein, P.C.: Effects of age on motor preparation and restructuring. *Bulletin of the Psychonomic Society* 27, 199–202 (1989)
2. Goggin, N.L., Stelmach, G.E.: Age-related differences in kinematic analysis of perceptual movements. *Canadian Journal on Aging* 9, 371–385 (1990)
3. Salthouse, T.A.: Reasoning and spatial abilities. In: Craik, F.I.M., Salthouse, T.A. (eds.) *The handbook of aging and cognition*, pp. 167–211. Erlbaum, Hillsdale (1992)
4. Kelly, C.L., Charness, N.: Issues in training older adults to use computers. *Behavioral and Information Technology* 14, 107–120 (1995)
5. Salthouse, T.A.: Steps towards the explanation of adult age differences in cognition. In: Perfect, T.J., Maylor, E.A. (eds.) *Models of cognitive aging*, pp. 19–50. Oxford University Press, New York (2000)
6. Jacob, R.J.K.: What you look at is what you get: Eyemovement- based interaction technique. In: *Proceedings of ACM CHI 1990*, pp. 11–18 (1990)
7. Jacob, R.J.K.: The use of eye movements in human-computer interaction techniques: What you look at is what you get. *ACM Transactions on Information Systems* 9, 152–169 (1991)

8. Jacob, R.J.K.: Eye-movement-based human-computer interaction techniques: Towards non-command interfaces. In: Harston, H.R., Hix, D. (eds.) *Advances in Human-Computer Interaction*, pp. 151–190. Ablex, Norwood (1993)
9. Jacob, R.J.K.: What you look at is what you get: Using eye movements as computer input. In: *Proceedings of Virtual Reality Systems 1993*, pp. 164–166 (1993)
10. Jacob, R.J.K.: Eye tracking in advanced interface design. In: Baefield, W., Furness, T. (eds.) *Advanced Interface Design and Virtual Environments*, pp. 212–231. Oxford University Press, Oxford (1994)
11. Jacob, R.J.K., Sibert, L.E., Mcfarlanes, D.C., Mullen, M.P.: Integrality and reparability of input devices. *ACM Transactions on Computer-Human Interaction*, 2–26 (1994)
12. Sibert, L.E., Jacob, R.J.K.: Evaluation of eye gaze interaction. In: *Proceedings of CHI 2000*, pp. 281–288 (2000)
13. Murata, A.: Eye-gaze input versus mouse: cursor control as a function of age. *International Journal of Human-Computer Interaction* 21, 1–14 (2006)
14. Murata, A., Miyake, T.: Effectiveness of Eye-gaze Input System – Identification of Conditions that Assures High Pointing Accuracy and Movement Directional Effect. In: *Proceedings of 4th International Workshop on Computational Intelligence & Applications*, pp. 127–132 (2008)
15. Murata, A., Moriwaka, M.: Effectiveness of the menu selection method for eye-gaze input system -Comparison between young and older adults. In: *Proceedings of International Workshop on Computational Intelligence and Applications, IW CIA 2009*, pp. 306–311 (2009)
16. Murata, A., Moriwaka, M.: Basic Study for Development of Web Browser suitable for Eye-gaze Input System-Identification of Optimal Click Method. In: *Proceedings of 5th International Workshop on Computational Intelligence & Applications*, pp. 302–305 (2009)

Proposal of Estimation Method of Stable Fixation Points for Eye-gaze Input Interface

Atsuo Murata, Takehito Hayami, and Keita Ochi

Graduate School of Natural Science and Technology,
Okayama University, Okayama, Japan
{murata,hayami}@iims.sys.okayama-u.ac.jp

Abstract. As almost all of existing eye-gaze input devices suffers from fine and frequent shaking of fixation points, an effective and stable estimation method of fixation points has been proposed so that the obtained stable fixation points enabled users to point even to a smaller target easily. An estimation algorithm was based on the image processing technique (Hough transformation). An experiment was carried out to verify the effectiveness of eye-gaze input system that made use of the proposed estimation method of fixation point. From both evaluation measures, the proposed method was found to assure more stable cursor movement than the traditional and commercial method.

Keywords: Eye-gaze input, fixation point, stabilization, task completion time, pointing error.

1 Introduction

The technology for measuring a user's visual line of gaze in real time has been advancing. Appropriate human-computer interaction techniques that incorporate eye movements into a human-computer dialogue has been developed [1-11]. These studies have found the advantage of eye-gaze input system. However, few studies except Murata [8] have examined the effectiveness of such systems with older adults. Murata [8] discussed the usability of an eye-gaze input system to aid interactions with computers for older adults. Systematically manipulating experimental conditions such as the movement distance, target size, and direction of movement, an eye-gaze input system was found to lead to faster pointing time as compared with mouse input especially for older adults. Eye-gaze input interfaces [1-11] are paid more and more attention as an alternative to a mouse especially for disabled persons. As the eye-gaze input interface enables users to operate PC by eye movements, even disables persons with deficiency on the upper limb can easily use it. However, at present, it is difficult to obtain a stable fixation points so that one can point to a smaller target using an eye-gaze input system.

Almost all of existing eye-gaze input devices suffers from fine and frequent shaking of fixation points. Since the edge of iris and pupil is changing smoothly during the

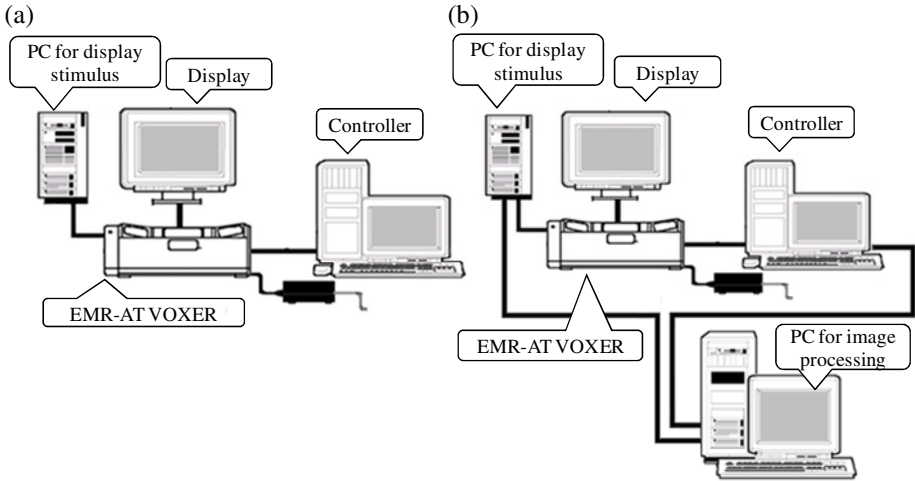


Fig. 1. (a) commercial eye-gaze measurement system used in this study and (b) addition of image processing system to the system in Fig.1(a)

extraction process of pupil image, it is difficult to extract the pupil image accurately and always obtain a stable coordinate of the pupil image. Due to this, an effective and stable estimation method of fixation points has not been established, and it is presently not easy to point to a smaller target such as ones used on GUI of Internet Explorer using the existing eye-gaze input system. An estimation algorithm based on the image processing technique (Hough transformation) had been proposed so that the obtained stable fixation points enabled users to point even to a smaller target easily.

The effectiveness of the proposed method for stably estimating fixation point and preventing the fixation points of the system from shaking was empirically verified. In the verification experiment, an easy pointing task using an eye-gaze input system was taken up. The task completion time, the operation error, and the cursor movement trajectory (the mean distance from the center of the target and the standard deviation of the coordinates) were compared between the traditional and the proposed methods.

2 Method for Estimating Stable Fixation Points

An estimation algorithm based on the image processing technique (Hough transformation) and detection of Purkinje image had been proposed so that the obtained stable fixation points enabled users to point even to a smaller target easily. The commercial eye-gaze measurement system used in this study is shown in Fig.1(a). The image processing system was added to the system (See Fig.1(b)).

When gazing at the lower part of the display, both eyelash and Purkinje image overlaps a pupil and consequently the pupil cannot be extracted. Due to this, stable eye fixation points cannot be obtained. In order to measure fixation points stably, we must manage to compensate for the lack of pupil image. In this study, we used Labview (NATIONAL INSTRUMENTS), added image processing component to the

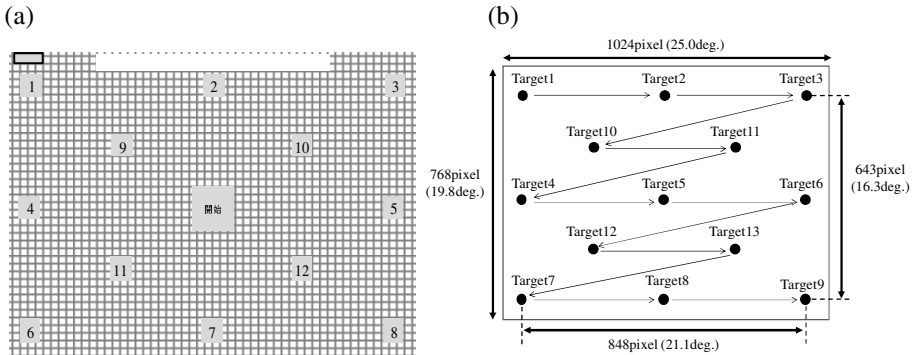


Fig. 2. (a) Experimental display and (b) Procedure of pointing task (From Target1 to Target9)

commercial system as in Fig.1(a), and made an attempt to extract pupil images using Hough transformation and obtain stable fixation points.

The pupil image was extracted by applying Hough transformation of a circle to an extracted edge as a candidate of pupil boundary. By setting the radius of the circle, Purkinje image can also be detected. It is well known that the movement of eye-gaze point is nearly proportional to the rotation of an eyeball. An attempt was made to estimate eye-fixation points on the basis of the change of distance between pupil and Purkinje image using the method by Mackworth et al. [12].

3 Experimental Method

3.1 Participants

Using five undergraduate or graduate students aged from 21 to 23, the usability of the proposed eye-gaze input system was experimentally compared with that of the conventional system. All participants were

3.2 Experimental Task

The participants were required to point to a target on a display as accurately and quickly as possible. The task completion time and the pointing accuracy were measured. The illumination and brightness on the experimental display were 500 lx and 98cd/m², respectively. The viewing distance was about 450mm.

Three kinds of squares (60 X 60 pixel² (2.09 degree of visual angle), 40 X 40 pixel² (1.39 degree of visual angle), and 30 X 30 pixel² (1.05 degrees of visual angle)) were used in the experiment. The participant was required to point to the targets in ascending order from Target1 to Target10 (See Fig.2(a) and (b)). The outline of experimental situation is shown in Fig.3.



Fig. 3. Outline of experimental situation

3.3 Design and Procedure

The cursor movement to the target was conducted using an eye-gaze system, and the click was done using a mouse when the cursor entered the target square. When the target was successfully pointed to, a beep sound rang and the movement to the next target started. The participant must judge whether the target was successfully clicked using only the click sound. When all of 10 targets were clicked, one task was completed.

4 Results

4.1 Task Completion Time

The task completion time is shown as a function of target size (30, 40, and 60 pixel square) and estimation method of fixation point (the commercial and the proposed ones) in Fig.4(a). When the target was 30 X 30 pixel², the mean task completion time of the proposed method was reduced by about 34% as compared with that of the traditional and commercial method. In Fig.4(b), the mean task completion time is plotted as a function of target location and estimation method of fixation point (the conventional and the proposed ones) in case of target size of 30 pixel.

4.2 Number of Errors

The number of errors is shown as a function of target size (30, 40, and 60 pixel square) and estimation method of fixation point (the commercial and the proposed ones) in Fig.5(a). In Fig.5(b), the mean task completion time is plotted as a function of target location and estimation method of fixation point (the conventional and the proposed ones) in case of target size of 30 pixel. The number of input errors for the proposed method was reduced by about 54% as compared with that of the traditional and commercial method. No differences of performance were detected between both methods when the target sizes were 40 X 40 pixel² and 60 X 60 pixel².

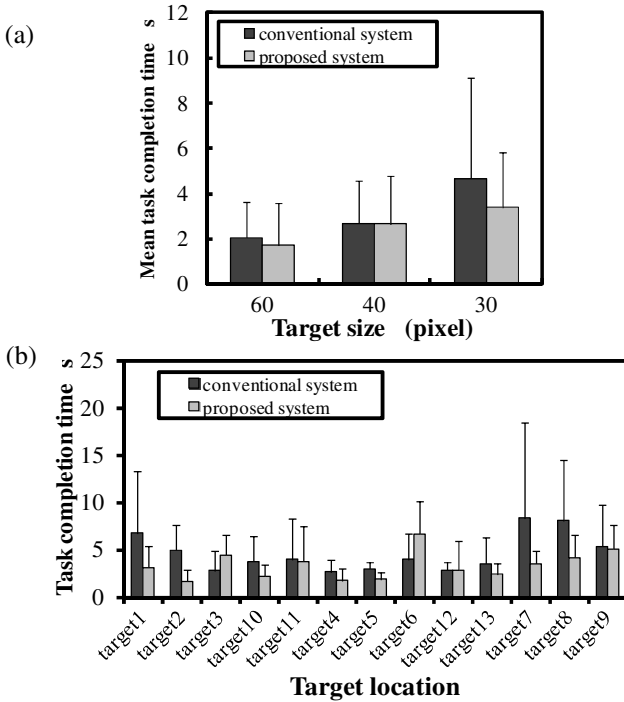


Fig. 4. (a) Mean task completion time as a function of target size (30, 40, and 60 pixel square) and estimation method of fixation point (the conventional and the proposed ones), (b) Mean task completion time as a function of target location and estimation method of fixation point (the conventional and the proposed ones) (target size: 30 pixel)

4.3 Cursor Movement Trajectory

The significant difference of performance (task completion time and number of errors) between two methods (traditional and proposed methods) was detected for the square target of 30 X 30 pixel². Therefore, the stability of cursor movement trajectory was analyzed only for the square target of 30 X 30 pixel². The cursor movement trajectory until 10 sample points (1/30 s X 10=1/3 s) before the mouse click was used for the analysis of cursor movement stability. Using these data, the mean distance from the center of the target and the standard deviation of the coordinates were calculated. The results for the target size of 30 pixel are depicted in Fig.6(a) and (b). From both evaluation measures, the proposed method was found to assure more stable cursor movement than the traditional and commercial method.

4.4 Performance for Target Size of 20 Pixel

In order to further verify the effectiveness of the proposed method, the data for the target size of 20 pixel was also collected. Fig.7(a) shows the mean task completion time as a function of estimation method of fixation point (the conventional and the

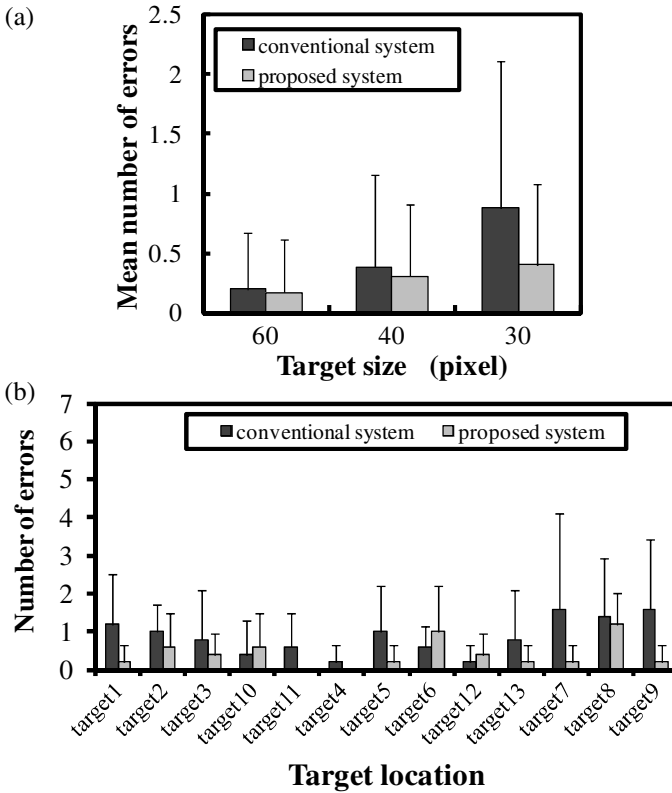


Fig. 5. (a) Mean number of errors as a function of target size (30, 40, and 60 pixel square) and estimation method of fixation point (the conventional and the proposed ones) and (b) Mean number of errors as a function of target location and estimation method of fixation point (the conventional and the proposed ones) (target size: 30 pixel)

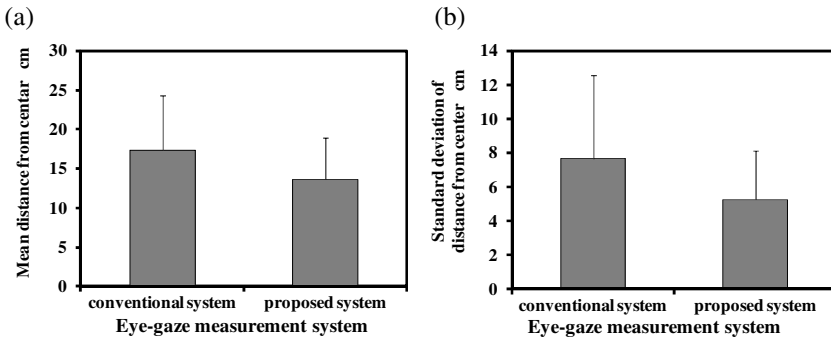


Fig. 6. (a) Mean distance from center as a function of estimation method of fixation point (the conventional and the proposed ones) (target size: 30 pixel) and (b) Standard deviation of distance from center as a function of estimation method of fixation point (the conventional and the proposed ones) (target size: 30 pixel)

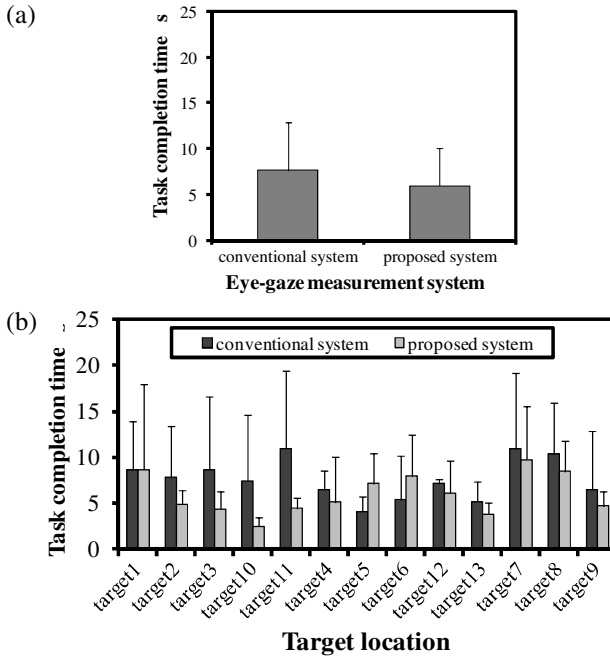


Fig. 7. (a) Mean task completion time as a function of estimation method of fixation point (the conventional and the proposed ones) (target size: 20 pixel) and (b) Mean task completion time as a function of target location and estimation method of fixation point (the conventional and the proposed ones) (target size: 20 pixel)

proposed ones) for the target size of 20 pixel. Fig.7(b) shows the mean task completion time as a function of target location and estimation method of fixation point (the conventional and the proposed ones) for the target size of 20 pixel. In Fig.8(a), the mean number of errors is shown as a function of estimation method of fixation point (the conventional and the proposed ones) for the target size of 20 pixel. In Fig.8(b), the mean number of errors is plotted as a function of target location and estimation method of fixation point (the conventional and the proposed ones) for the target size of 20 pixel. Fig.9(a) shows the mean distance from center as a function of estimation method of fixation point (the conventional and the proposed ones) for the target size of 20 pixel. Fig.9(b) compares the standard deviation of distance from center between the two estimation methods of fixation point (the conventional and the proposed ones) for the target size of 20 pixel.

5 Discussion

5.1 Task Completion Time

As shown in Fig.5(a) and Fig.7(a), it was confirmed that the task completion time of the proposed system tended to be shorter as compared with that of the traditional

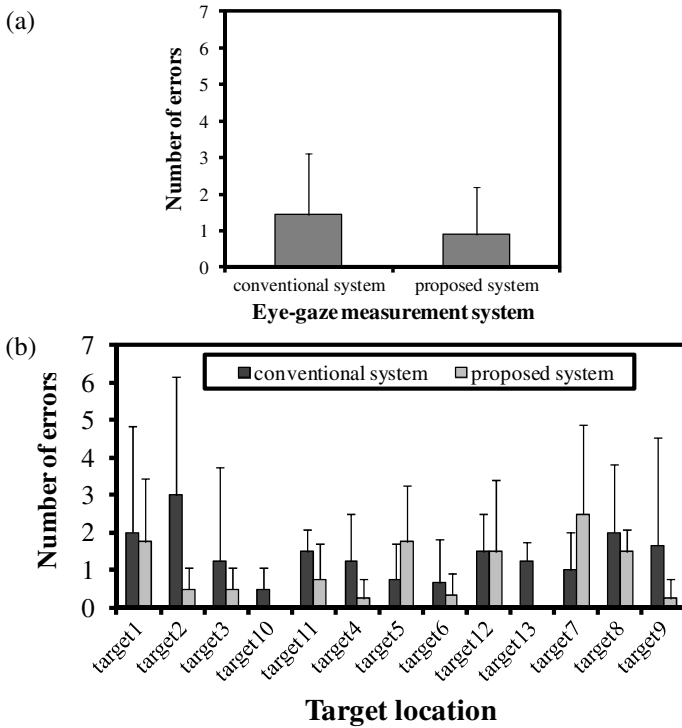


Fig. 8. (a) Mean number of errors as a function of estimation method of fixation point (the conventional and the proposed ones) (target size: 20 pixel) and (b) Mean number of errors as a function of target location and estimation method of fixation point (the conventional and the proposed ones) (target size: 20 pixel)

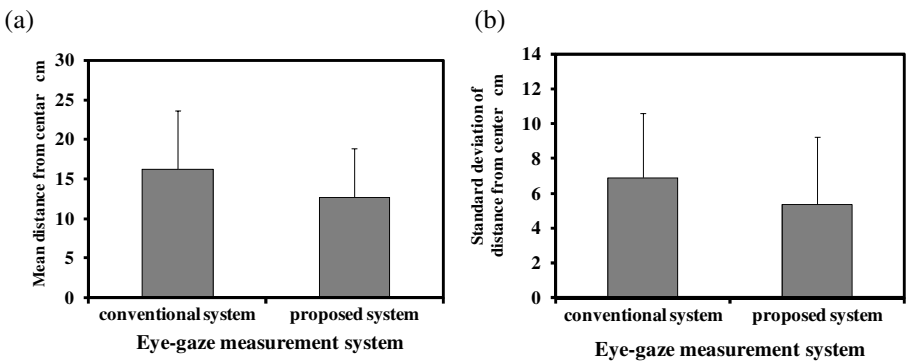


Fig. 9. (a) Mean distance from center as a function of estimation method of fixation point (the conventional and the proposed ones) (target size: 20 pixel) and Standard deviation of distance from center as a function of estimation method of fixation point (the conventional and the proposed ones) (target size: 20 pixel)

system especially when the target size was smaller (20 pixel and 30 pixel). When the cursor shakes to a larger extent, it is difficult to keep the cursor inside the smaller target (20 or 30 pixel), and consequently the task completion time is prolonged. When the target size is large enough relative to the shaking of the cursor, the shaking of the cursor does not affect the task completion time. Therefore, no significant difference of task completion time must be detected between the traditional and the proposed methods when the target size was 60 pixel or 40 pixel as shown in Fig.5(a). The task completion time of the proposed system for the target size of 20 pixel was reduced by 36% as compared with that of the traditional system (see Fig.7(a)). The proposed system is more effective when the target size is smaller and less than 30 pixel.

5.2 Mean Number of Errors

As shown in Fig.6(a) and Fig.8(a), the error trial tended to be less for the proposed method than for the traditional method in particular when the target size was less than 30 pixel. In case of the target size of 30 pixel, the number of error for the proposed method was reduced by 54% as compared with that of the traditional system. When the target size was 20 pixel, the number of error for the proposed method was reduced by 44% as compared with that of the traditional system. It tended that the error pointing frequently occurs at the target location 7, 8, and 9 in Fig.2(b) when the traditional method was used. The following problems in the traditional method must be due to such a frequent error at the lower part of the display. When viewing the lower part of the display, the pupil image is to a larger extent covered by the eyelash or Purkinje image. Such a problem could be overcome by the proposed method, which led to the stable estimation of fixation points. The error data also supported the effectiveness of the proposed method.

5.3 Cursor Movement Trajectory

The participants must click the target using an eye-gaze input system at the instance when the participants felt that the click could be properly carried out. Therefore, the cursor movement trajectory before and after the click operation was examined to check whether the cursor movement was stable or not.

From Fig.6(a) and (b) and Fig.9(a) and (b), it is clear that the cursor position of the proposed method was nearer to the center of the target and less dispersive than that of the traditional method. This means that the reduced shaking of the cursor by the proposed method led to the nearer click to the center of the target, and less dispersive click location.

5.4 Implication for HCI Design

When the target size was large (more than 40 pixel), no significant differences of the task completion time and the number of errors were detected between the conventional and the proposed methods. Therefore, both methods are applicable to larger target more than 40 pixel. When the target is less than 30 pixel, the proposed method should

be used, because the proposed method is superior to the conventional method from the perspectives of task completion time, the entry error, and the stability of clicked coordinates.

The validity of the present study should be verified by increasing the number of participants. Future work should explore whether more stable fixation point can be obtained by adding the smoothing technology of coordinates to the proposed method.

References

1. Jacob, R.J.K.: What you look at is what you get: Eyemovement- based interaction technique. In: Proceedings of ACM CHI 1990, pp. 11–18 (1990)
2. Jacob, R.J.K.: The use of eye movements in human-computer interaction techniques: What you look at is what you get. *ACM Transactions on Information Systems* 9, 152–169 (1991)
3. Jacob, R.J.K.: Eye-movement-based human-computer interaction techniques: Towards non-command interfaces. In: Harston, H.R., Hix, D. (eds.) *Advances in Human-Computer Interaction*, pp. 151–190. Ablex, Norwood (1993)
4. Jacob, R.J.K.: What you look at is what you get: Using eye movements as computer input. In: Proceedings of Virtual Reality Systems 1993, pp. 164–166 (1993)
5. Jacob, R.J.K.: Eye tracking in advanced interface design. In: Baefield, W., Furness, T. (eds.) *Advanced Interface Design and Virtual Environments*, pp. 212–231. Oxford University Press, Oxford (1994)
6. Jacob, R.J.K., Sibert, L.E., Mcfarlanes, D.C., Mullen, M.P.: Integrality and reparability of input devices. *ACM Transactions on Computer-Human Interaction*, 2–26 (1994)
7. Sibert, L.E., Jacob, R.J.K.: Evaluation of eye gaze interaction. In: Proceedings of CHI 2000, pp. 281–288 (2000)
8. Murata, A.: Eye-gaze input versus mouse: cursor control as a function of age. *International Journal of Human-Computer Interaction* 21, 1–14 (2006)
9. Murata, A., Miyake, T.: Effectiveness of Eye-gaze Input System -Identification of Conditions that Assures High Pointing Accuracy and Movement Directional Effect. In: Proceedings of 4th International Workshop on Computational Intelligence & Applications, vol. 127-132 (2008)
10. Murata, A., Moriwaka, M.: Effectiveness of the menu selection method for eye-gaze input system-Comparison between young and older adults. In: Proceedings of International Workshop on Computational Intelligence and Applications (IWCIA 2009), pp. 306–311 (2009)
11. Murata, A., Moriwaka, M.: Basic Study for Development of Web Browser suitable for Eye-gaze Input System - Identification of Optimal Click Method. In: Proceedings of 5th International Workshop on Computational Intelligence & Applications, pp. 302–305 (2009)
12. Mackworth, J.F., Mackworth, N.H.: Eye fixations recorded on changing visual scenes by the television eye-marker. *Journal of the Optical Society of America* 48, 439–445 (1958)

Modeling Situation-Dependent Nonverbal Expressions for a Pair of Embodied Agent in a Dialogue Based on Conversations in TV Programs

Keita Okuuchi^{1*}, Koh Kakusho¹, Takatsugu Kojima², and Daisuke Katagami³

¹ Kwansei Gakuin University, Sanda, Japan
{ccd77074, kakusho}@kwansei.ac.jp

² Shiga University of Medical Science, Otsu, Japan

³ Tokyo Polytechnic University, Tokyo, Japan

Abstract. Mathematical model for controlling nonverbal expressions of a pair of embodied agents designed for presenting various information through their dialogue is discussed. Nonverbal expressions of a human during conversation with others depend on those of them as well as the situation of the conversation. The proposed model represents the relationship between nonverbal expressions of a pair of embodied agents in different situations of conversation by a constraint function, so that the nonverbal expression of each agent reproduces the characteristic of nonverbal expressions observed in human conversation with various situations in TV programs by minimizing the function.

Keywords: Embodied agent, Human-agent interaction, Nonverbal expression.

1 Introduction

It is proposed in some recent work on human-agent interaction to employ a dialogue by a pair of embodied agents for presenting information to users, similar to news or talk shows in TV[1][2]. Information presentation based on a dialogue is expected to make the point of the information easier to understand than that based on a monologue.

For realizing dialogues between embodied agents with nonverbal expressions, we need to consider how to maintain consistency between those nonverbal expressions, such as each agent should nod and smile when the other agent speaks with a smile. Aiming to maintain the consistency, we have proposed a mathematical model that represents the consistency by referring to social psychological studies about the relationship between nonverbal expressions of humans in a dialogue[3]. However, the degree of each nonverbal expression is not always the same but could differ with changing situations in the dialogue.

In this article, we discuss which kinds of situations actually affect the degree of nonverbal expressions in what manner, by analyzing human dialogues in TV programs, in order to propose a further extension of the mathematical model, which approximates the dependency of nonverbal expressions on those kinds of situations.

* Corresponding author.

2 Dependency of Nonverbal Expressions on Changing Situations in a Dialogue

2.1 Relationship between Nonverbal Expressions

In the previous work on social psychology, it is known that positive correlations are observed between the amount of speech and the degree of nonverbal expressions, which include gaze, smile and nod, of humans during conversation; when the amount of speech or the degree of one of the nonverbal expressions increases or decreases, the degree of another nonverbal expression increases or decreases in a similar manner[4][5][6][7]. These positive correlations are summarized in Table 1, where each pair denoted by #1-12 is reported to show the positive correlation for their amounts in human conversation as described above.

Table 1. Correlations of nonverbal expressions

(a) Between different persons

		PersonA			
		speech	gaze	smile	nod
PersonB	speech		#4	#6	#8
	gaze	#1	#5		
	smile	#2		#7	
	nod	#3			

(b) Within the same person

	gaze	nod	smile	hand gesture
speech	#9	#10	#11	#12

2.2 Factors for Classifying Situations of Human Conversation in TV Programs

Since the above report of the previous work only describes the general tendency in the appearance of nonverbal expressions, the degree of each nonverbal expression in actual conversation could differ with various kinds of situations changing during the same dialogue while satisfying those correlations qualitatively. In order to learn about most influential situations especially in dialogues for information presentation, we made some interviews for 18 participants.

The participants are asked to watch some news and talk shows in TV while paying their attentions to the moments when they recognize the situation of the conversation changes. For each moment that they recognize the change, they are asked about the factor that cause them to recognize the change.

The result is shown in Figure 1. In this figure, the number of answers of the participants is counted by classifying the answers into three kinds: (A)role of the interlocutors (which interlocutor takes the control of the conversation), (B)structure of participation (which are the addressee of the speech by each interlocutor, the other interlocutor or the viewer), (C)atmosphere of the dialogue (humorous or serious). These three factors

are based on the viewpoint taken in the study of conversation analysis[8][9]. This result shows that most of the factors that make the participants recognize the change of situations are included in one of these three factors.

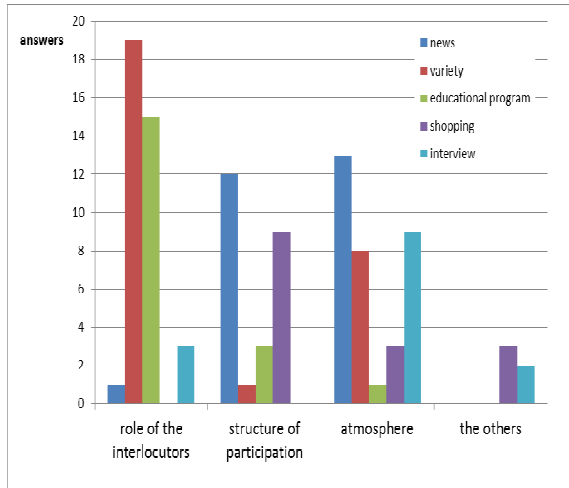


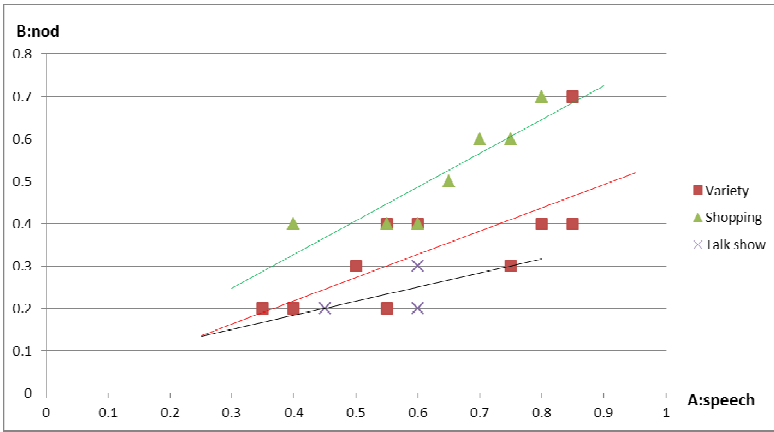
Fig. 1. Factors for recognizing different situations of TV programs

2.3 Analyzing the Dependency of Nonverbal Expressions on the Situations of Conversation

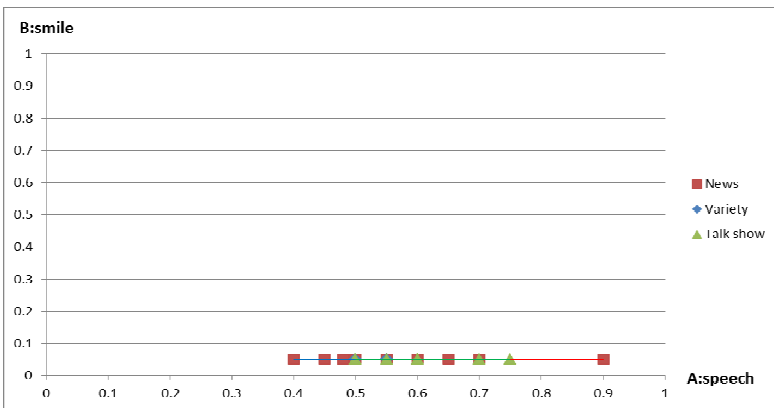
In order to analyze the dependency of the positive correlations between nonverbal expressions on situations (A)-(C) described in 2.2, we observed actual degree of each nonverbal expression in dialogues of various TV programs. If the amount of speech or the degree of a nonverbal expression of one person has positive correlation with those of the other person, the positive correlation should be approximated by a linear function describing the rate of the amount or the degrees between the positively correlated pair of speech or nonverbal expressions, and if the correlation depends on the situation of conversation, the rate should take different values for different situations.

From the result of our analysis about how each pair of the amount of speech or the degrees of nonverbal expressions, we found that some pairs can be approximated by linear functions and others cannot. Figure 2 are examples of the pairs that can be approximated by linear functions and those that cannot. For example, the degree of nod of interlocutor B has the positive correlation that can be approximated by a linear function with the amount of speech of interlocutor A for the situation of conversation with the structure of participation where the addressee of each interlocutor is the other interlocutor in humorous atmosphere, whereas the degree of smile of interlocutor B has no relation with the amount of speech of interlocutor A for the structure of participation where the addressee of each interlocutor is the other interlocutor in serious atmosphere. The pair of the amount of speech or the degrees of nonverbal expressions with positive

correlation for each situation of conversation is summarized in Table.2, where each cell with a check denotes that the corresponding pair has positive correlation.



(a) With positive correlation
("humorous" & "addressing to the other interlocutor")



(b) Without positive correlation.
("serious" & "addressing to the other interlocutor")

Fig. 2. Example of a pair of speech or nonverbal expressions with or without positive correlation for their amount or degrees

Figure 2(a) also shows, from the slopes of the lines, that the same pair of the amount of speech or the degrees of nonverbal expressions takes different ratios for their positive correlation relationship for different kinds of programs. For example, the degree of nod of interlocutor B increases more rapidly with the increase of the amount of speech of interlocutor A in shopping programs than in variety shows or talk shows for the situation of conversation with the structure of participation where the addressee of each interlocutor is the other interlocutor in humorous atmosphere.

Table 2. Correlation between nonverbal expressions in different situations

situation	I	II	III	IV
(B)addressee	intolocutors	intolocutors	viewers	viewers
(C)atmosphere	humorous	serious	humorous	serious
#1	✓	✓		
#2	✓		✓	
#3	✓	✓	✓	✓
#4	✓	✓		
#5	✓	✓	✓	✓
#6	✓		✓	
#7	✓	✓	✓	✓
#8	✓	✓	✓	✓
#9	✓	✓		
#10	✓	✓	✓	✓
#11	✓		✓	
#12	✓	✓	✓	✓

3 Mathematical Model for Situation-Dependent Correlations

In order to reproduce the situation-dependent positive correlation between nonverbal expressions in TV programs described in section 2 by a pair of embodied agent in a dialogue following a predetermined scenario for information presentation, we represent the positive correlation by a constraint function.

First of all, we represent the basic relationship of positive correlation in the amount of speech or the degrees of nonverbal expressions for pairs #1-12 by constraint function E as follows:

$$E \equiv \sum_{i=1}^{16} E_i = \sum_{i=1}^{16} (x_i^X - y_i^Y) = 0 \tag{1}$$

where x_i^X and y_i^Y denote the amount of speech or the degrees of the nonverbal expressions for pair #i (i=1,...,12) to be produced by embodied agents X and Y, which are variables to denote one of the two embodied agents A and B ($X, Y \in \{A, B\}$).

As already shown in Table 2, it depends on the situation of communication whether the positive correlation is actually observed or not for each pair #1-12. In order to represent the situation-dependency for the existence of positive correlation for each pair #1-12, we modify equation (1) by introducing an additional variable that represents the existence of positive correlation for each pair as follows:

$$E' \equiv \sum_{i=1}^{16} l_i E_i = \sum_{i=1}^{16} l_i (x_i^X - y_i^Y) = 0 \tag{2}$$

where variable l_i denotes the binary flag that represents the existence of the positive correlation between the amount of speech or the degrees of nonverbal expressions denoted by pair #i. This value is predetermined based on Table 2.

Moreover, as already discussed in section 2.2, the ratio between the values for the amount of speech or the degrees of nonverbal expressions depends on the kind of TV program. In order to represent this dependency, we modify equation (2) by introducing another additional variable that represents the ratio as follows:

$$E'' \equiv \sum_{i=1}^{16} E_i' = \sum_{i=1}^{16} (x_i^X - \alpha_i y_i^Y) = 0 \tag{3}$$

where variable α_i denotes the ratio between the value for the amount of speech or the degrees of nonverbal expressions for pair #i. This value is estimated in advance so that function E'' is minimized when the degrees of nonverbal expressions observed in each situation of actual TV programs are given as the value of x_i^X and y_i^Y .

For controlling nonverbal expressions of embodied agents A and B in each situation of conversation, the values of x_i^X and y_i^Y for each moment of the dialogue are calculated by minimizing function E'' by setting the degree of the speech of each embodied agent at the moment based on the scenario of the dialogue, after the values of l_i and α_i for the situation are obtained as described above.

4 Experimental Results

The degrees of nonverbal expressions is calculated for different situation of conversation by minimizing constraint function E'' in equation (3) after setting the amount of speech of interlocutors of actual TV program used in section 2 for the values of variables x_i^X and y_i^Y corresponding to speech. The resultant values for the degrees of nonverbal expressions for a variety show are shown in Figure 3. As shown in the figure, the tendency for degrees of nonverbal expressions changes with situations of communication I – III in Table.2. The reason why the degrees of each nonverbal expression are not the same even in the same situation is because it is also affected by the amount of speech at each moment.

In order to realize a dialogue by a pair of embodied agents with speech and nonverbal expressions obtained above, we employed TVML (TV program Making Language)[10], which is developed by NHK (Japan Broadcasting Cooperation) for producing TV program by a script language. Sample scenes in situations I - III are shown in Figure 4. As shown in the figure, the degrees of nonverbal expressions of each embodied agent are controlled depending on changing situations of the dialogue. For example, the agents gaze at each other in the situation where the structure of participation where the addressee of each agent is the other agent (situation I, II), whereas they gaze at the viewer in the situation where the addressee is the viewer (situation III). Also, the smile of the agents increases for humorous atmosphere (situation I, III), whereas it becomes zero for serious atmosphere (situation II). In addition, the

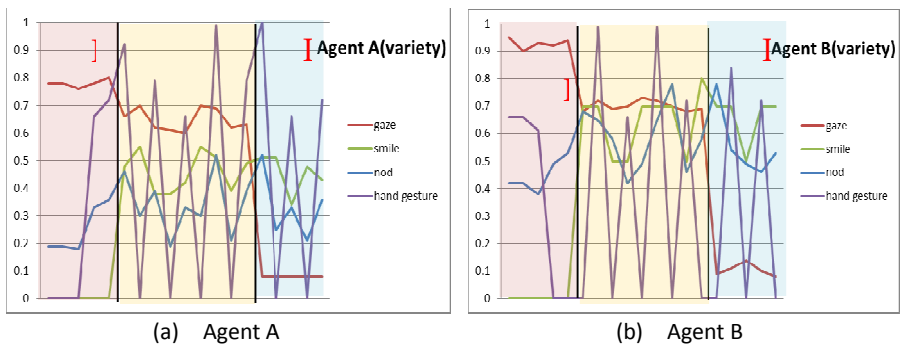


Fig. 3. The degrees of nonverbal expression reproduced by the proposed model for different situations of communication

agents do not gaze at each other all the time but also gaze sometimes at the viewer even in the situation with the structure of participation where the addressee of each agent is the other agent (situation I).

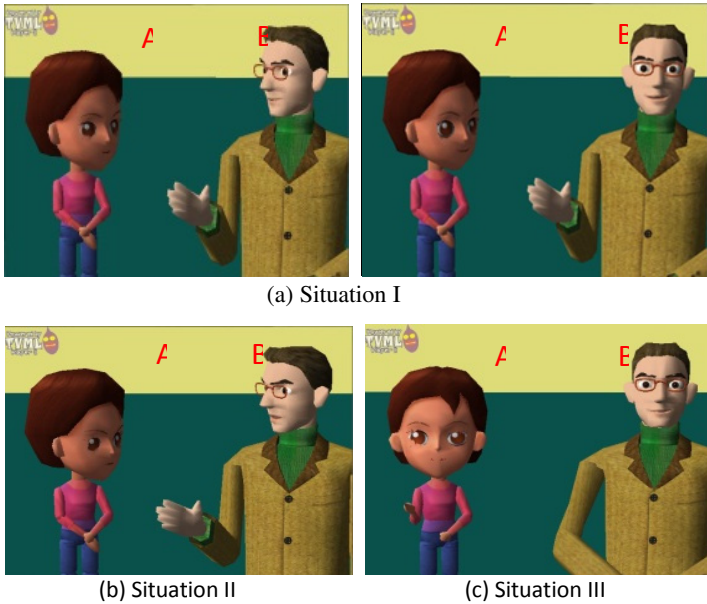


Fig. 4. Sample scenes of the dialogue by the embodied agents in different situations

5 Conclusion

We proposed a mathematical model that approximates situation-dependency of positive correlation between nonverbal expressions in human conversation in TV program for a pair of embodied agent in a dialogue for information presentation to users. As shown in the experimental results, our model reproduces the characteristic of nonverbal expressions in human conversation depending on the situation of the structure of participation and atmosphere. However, for increasing the precision for reproducing the characteristic of nonverbal expression in human conversation, we need to consider further improvement of our model as one of our future steps.

In our model, the degree of nonverbal expressions of embodied agents are driven by the speech, which is given as the script for their dialogue, based on its constraint of on the nonverbal expressions represented by the constraint function. This constraint only employs the amount of speech without consideration of the content of the speech. However, the nonverbal expression in human conversation should also depend on the content of the speech. In order to extend our model towards the capability for considering the dependency of nonverbal expression on the content of the speech, we plan to include some tags that represent various situations dependent on the content of the speech in the script of the dialogue of the agents.

References

1. Kubota, H., Yamashita, K., Nishida, T.: Conversational Contents Making a Comment Automatically. In: Int. Conf. on Knowledge-Based Intelligent Information Engineering Systems & Allied Technologies (KES), pp. 1326–1330 (2002)
2. Takahashi, T., Katagami, D.: Agent Design As An Information Interface Assumed For Full-time Operation. In: Human-Agent Interaction Symposium (2011) (in Japanese)
3. Kakusho, K., Itou, J., Minoh, M.: An Embodied Agent that Sends Nonverbal Conversational Signals Consistent with those of the Partner during a Dialogue. In: IEEE Workshop on Robot and Human Interactive Communication (RO-MAN), pp. 247–252 (2003)
4. Beattie, G.W.: Sequential Temporal Patterns of Speech and Gaze in Dialogue, *Semiotica* (1978)
5. Dimberg, U.: Facial Reactions to Facial Expressions, *psychophysiology. psychophysiology* 19(6), 643–647 (1982)
6. Kendon, A.: Some functions of gaze direction in social interaction. *Acta Psychologica* 26, 22–63 (1967)
7. Matarazzo, J.D., Saslow, G., Wiens, A.N., Weitman, M., Allen, B.V.: Interviewer Head Nodding and Interviewee Speech Durations. *Psychotherapy: Theory, Research and Practice* 1, 54–63 (1964)
8. Clark, H.H.: *Using Language*. Cambridge University Press (1996)
9. Gatica-Perez, D.: Automatic nonverbal analysis of social interaction in small groups. *Journal Image and Vision Computing* 27(12) (2009)
10. <http://www.nhk.or.jp/str1/tvml/>

Research on a Large Digital Desktop Integrated in a Traditional Environment for Informal Collaboration

Mariano Perez Pelaez, Ryo Suzuki, and Ikuro Choh

Waseda University, Department of Intermedia Art and Science, Japan
mperez@aoni.waseda.jp, reputless@gmail.com, choh@waseda.jp

Abstract. We are building a digital desktop system designed to support the tasks that are usually performed around the traditional desktop. Tabletop platforms are not new environments, especially as a research topic, but most of the existent systems try to adapt the computer work style or only serve as platform for experimenting with new features. In contrast our targets are to support the traditional work flow around desktops, not forcing the users to modify their methods and to build the system as a complete tool for everyday tasks. We want to provide a usable environment with computer-support features for raising productivity and enhancing the user experience. For doing this we realized a field study about the traditional desktop activities and with this knowledge we designed new tools and features that fit the user real needs and environment.

Keywords: Natural interface, interaction design, workgroup support, collaborative environment.

1 Introduction

1.1 Research Outline

We consider as a traditional desktop a large table in which many users can get together to perform a task and to collaborate. Desktops are placed everywhere, from schools to offices, and therefore are common places for informal collaboration, creation of documents, brainstorming, meetings and a large variety of activities that involve different users working together and creating or sharing information. These activities are susceptible to be enhanced by computer support with the system we propose. We define a digital desktop as a computer-based desktop or tabletop capable of detecting the interaction of users on it (usually their fingers or other tools) and project information on the surface. Our system prototype is basically a wide digital desktop.

We are building a digital desktop system for collaboration and automation of activities performed on traditional desktops, especially those involving workgroup. As the base of our system, two desktops provide a very large workspace in which we project interactive digital information that can be created or manipulated by the users. We are focused in providing the users a work environment for performing everyday tasks, with support features in order to improve the productivity, the quality of the resulted contents and for a better user experience in global.

The current version of our desktop environment is the result of an evolving tabletop system that we are developing in our laboratory [1][2][12]. The project started as a tool for document creation but now has become more open system for any kind of tasks related to the desktop.



Fig. 1. The prototype of the digital desktop

1.2 Research Targets

From our experience in previous digital desktop researches, we decided the following targets and minimum requirements for our platform:

- Integrated with a traditional environment and non-intrusive. When working on the desktop we use different tools as notebooks, pens, laptop computers, etc. The system should allow the users to work with their traditional tools and with the desktop system at the same time.
- With a natural interaction model and a minimal interface. We do not want to replicate the traditional computer WIMP model or any other, but to design a simple interaction model that does not interfere with the work on the desktop. In other words, we want to provide to the users an interaction model as near as possible to the real desktop.
- The system should be a platform for collaboration. The system should allow groups of users to work in the desktop at the same time and the easy sharing of information between them.
- Computer based support features. With the previous targets we only had a system with the same functions of a traditional desktop. But we want to add new computer-supported features and tools to enhance the work on the desktop in different fields. These features will include automation of some tasks.

- Useful in different situations. As the activities performed on a desktop are very different (from working alone to a group meetings), we want the system to support as many situations as possible.
- External connectivity. The digital desktop will allow other devices to connect with the desktop and import/export/manipulate the information on the system.
- Remote location users support. As a collaborative platform, the digital desktop should also let remote users to collaborate with the users at the desktop in real time or asynchronously.

In consonance with these targets, to provide solid support to the users real needs and to respect their work process, we decided to perform a field study on traditional desktops environments described in chapter 2.

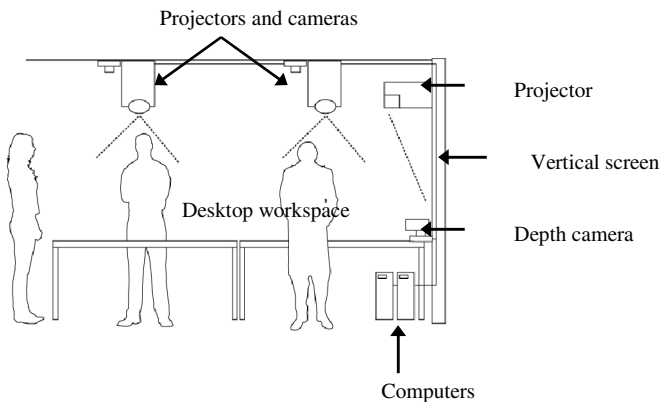


Fig. 2. Prototype structure

1.3 Background and Previous Researches

One of the first researches on digital desktop is the DigitalDesk [4], which introduces a desktop with projected information or applications and with the possibility of physical document text recognition for interacting with these applications. Other researches explored the desktop as a meeting and workgroup place for support co-located or remote collaboration [6][7] or add new features as document management [10].

Systems as SketchSpace [3] or ReacTable [8] explored ways of interacting with physical objects as part of a tabletop. Reactable used blocks that can be linked in a digital desktop environment to perform music. The research of Fiebrink, “Ensemble” [4], mixes physical devices in an interactive tabletop for sound edition, discovering that the users prefer to use physical devices rather than those based on a touch-capable desktop. Other researches as [11], which use physical tools similar to pencils, cutters and magnifiers for painting, obtained good evaluation by the users. From these researches we discover that users will prefer a physical tools depending of task they are performing.

Hartman [5] also studied how to introduce traditional input devices (keyboard and mouse) inside an interactive workgroup desktop environment. The researchers of SLAP widgets [14] proposed a set of physical objects for interact with a digital tabletop. Some of this objects are similar to traditional computers elements (buttons and keyboards) while other are more specially designed for the tabletop, as a knob that can be moved and rotated for selecting options or values in a menu.

The system described in [9] also uses a tabletop environment combined with personal devices forming a common workspace for remote collaboration oriented to a creative task.

2 Field Research

2.1 Studying the Work Process

We observed the use of traditional desktops inside the university environment (meeting room and group workspaces) for a period of 6 months. In order to support as many activities as possible, we covered different work situations and environments.

We specially focused on which actions and how are performed for making a task analysis in order to divide each task in simple atomic steps that we can enhance or automate. We also pay attention to the elements that are manipulated in the workspace and are susceptible to be integrated in the system.

Our target was also to differentiate the case when an interaction is with physical objects (traditional paper, notebooks, post-it, etc.), with digital information (using a laptop computer, tablet, etc. through a keyboard, mouse or other interface), or when both digital and physical are used together.

Together with the environment observation, we also utilized a questionnaire responded by 17 graduate and post-graduate students and individual interviews.

2.2 Results of the Study

When studying the users around the desktop, we discover very different work styles and processes. We discovered that 90% of the users feel useful their current digital and physical tools but only a lower percentage, 33%, is also satisfied with its current remote collaboration tools (videoconference, chat, e-mail, etc.). The tools they used are very different, including from simple blank paper to advanced laptop computers.

We also discovered, for example, that for sharing digital information between students that are working together, they print the document, creating a copy for every participant (even when a digital version is already available to all of them) or at least one hard copy in order to easily work on the same document while collaborating. We also discover other facts as they usually take a picture of the workspace when a task spans more than a session, in order to work individually or to recover the workspace to continue later.

With these results we confirmed or refined our targets for integrating the existent desktop elements in the digital environment and to support collaborative work and the

sharing of information. We specially designed tools and features to support the most common actions we observed, as shown on Table 1.

Table 1. Users most common actions and how are reflected in our system

Action	Original action	System approach
Creating a hand drawn picture or write a short text	physical	Pen tools or connected device
Sharing a element of physical information	physical	Scan tool with Clone tool
Precise manipulation	digital	Pointing device of a connected device
Sharing a element of digital information	digital	Clone tool
Working on the same physical element by a group	physical	Clone Tool or personal device access
Working on the same digital element by a group	digital	Clone Tool or personal device access
Writing / editing long texts	digital	Personal device with keyboard input
Workspace save/restoration	digital and physical	Automated
Document layout creation	digital and physical	Automated
Searching or reviewing old information	digital and physical	Recovery tool or review tool
Searching for reusing previous work	digital and physical	Review tool and recovery Tool or Clone tool

3 The System

3.1 System Description

At a glance the environment looks like a simple traditional desktop with a vertical whiteboard. Actually the system can be used completely as an "analog" desktop, using the digital extensions only when needed.

The system workspace consists of two large desktops with three depth cameras, two traditional cameras and two computers for managing the information (Fig. 2). The depth cameras manage the user interaction while the traditional cameras are used for image acquisition. The digital information is projected on the desktops and on the whiteboard using three projectors located on top of the system.

The touch recognition algorithm we use let the users to utilize the environment as a traditional desktop placing their laptop, notebooks or any other personal item on it without interfering with the system digital features.

3.2 Information on the System

The elements of the digital desktop can be of three different basic types: image, text or groups of elements. The users can manipulate freely these elements by direct touch

interaction. New elements can be added by grouping elements, uploading a file, downloading it from the Internet or from a connected device as described in 3.4.

This high simplicity of elements can be an inconvenient for complex tasks or those which need specific information types, but reduce the overhead of information a user needs to learn for utilize the system. By this principle we also do not use gestures to interact with these elements, as any other than moving has a low user evaluation [13].

In order to support specific situations, the system allows extending these basic types. As extensions we already developed a movie and a web browser types. These extensions are only activated when needed in an activity.

Also with the elements groups the system only control that certain elements are grouped, but it is possible to extend the system default with interpreters, giving meaning to groups and changing the way they are displayed. For example, a common group type is the “page” type, and by grouping pages we can create a “book” type that displays only two pages navigated with a flip animation.

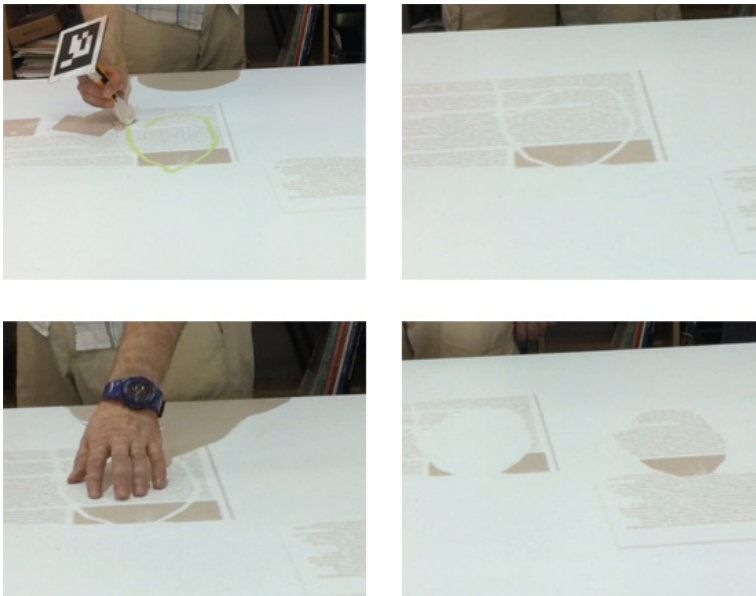


Fig. 3. Using a physical tool, a cutter, with digital information

3.3 Tools Based on Physical Counterparts

Touch interaction is a natural form to manipulate digital information, especially for moving and grouping elements, but is not adequate to other tasks performed in the environment. Depending on the situation, a more precise or specific input method is better than direct touch, as in the case of complex drawings, precise operations or manipulating a huge number of digital elements.

For some of these situations we propose the use of physical tools that are specifically well designed to perform certain tasks with physical elements to perform

the same or similar role with digital elements. When using these tools, our system recognizes the arms and hands that are holding them, in order to not interfere with the tool operation.

We first provided a set of colored pens to draw free hand on the system. The use is straightforward: take a color pen and draw directly on the desktop exactly as when using a pen on paper. It is possible to use the colored pen tools with other traditional tools as, for example, a ruler to draw straight lines. The same approach is used for the rest of tools of this category. For example we provide a cutter which can be used to divide digital images (Fig. 3) or documents, an eraser to delete information from the desktop, or a clone tool (in the form of a stamp) that can duplicate elements.

3.4 External Devices and Online Connectivity

Our tabletop system is not the only digital device that is used on the desktop. The users usually bring with them their personal devices as smartphones, tablets and laptop computers as valuable tools (almost 100% of the users bring any kind of digital device to the desktop and more than 50% is satisfied with them).

In order to connect the system with external devices the users only need a web browser and to access a special web site created by the system, which give interactive access to a display of the desktop. The connection, in real time, not only allows the sharing of information between devices, but also is very useful to overcome some limitations of the environment. We can, for example, take advantage of the high resolution of a tablet to read more conveniently documents or use the keyboard of a laptop computer to input text without the need of dedicated devices. Different users can access and read the same document in their own devices, even if the digital copy on the desktop is not physically near of them, or if they are outside the environment.

3.5 Automation and Other Features

The system provides different ways of automate repetitive tasks. These features include the automated sharing of the desktop and the continuous storing and restoring of its contents. These stored contents can also be displayed as a timeline to better visualization of the work process or to extract old information.

Other automation functionality includes the automated layout generation. Using a set of templates based on grid design and group of elements, the system can generate variation of the layout applied to the elements automatically.

For sharing physical information the system provides a tool for scanning the contents of the desktop. The system performs the scan instantly using one of the high-resolution cameras located on top of the environment, producing a digital copy automatically. If desired, the system can also process the image with an OCR algorithm to generate a digital text copy. This easy to use tool is a key point to break the barrier between physical and digital information without interrupting the workflow.

4 Evaluation

For evaluating the system we are performing two different types of experiments. First we are evaluating each of the features independently after finishing the implementation to test its usability. Each tool was evaluated by groups of three users receiving mixed reactions. For example, the color pen tool received a neutral acceptance, while the eraser and the cutter are considered “fun to use”.

The second type of evaluation takes small groups of 3 users (Fig. 4) for testing a wider set of features in simple tasks. Especially we are focusing in how they work on the environment and how they collaborate when co-located and when one user is in a distant location. We also ask the users to perform a similar task without using the desktop in order to compare the results and users preferences. The experiments till now show that these tasks took more time with the digital desktop than when working with a traditional desktop, but all the users feel that our system was useful and would use it in the future. In this stage, the difference of time can also be caused by the lack familiarization with the system features.

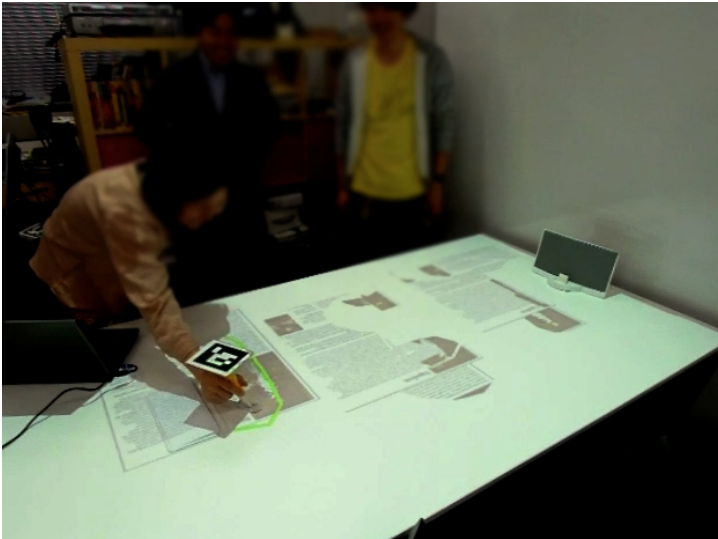


Fig. 4. Three users working in the desktop during an experiment.

5 Conclusion and Future Directions

We are building an open and multipurpose large desktop system to support different situations involving collaboration. While testing it we become aware of some issues that affected very negatively the integration of the system with the users existent work style. Though the hands and objects detection has been highly improved in the current version of the system, there are still recognition failures, especially when there is an

obstruction in the cameras field of view. Due to the system layout, this is very common when a user is standing in certain position or by inclined laptops screens, creating a shadow for the tracking system.

In the evaluation part we still need to evaluate the performance of a long time “real world” tasks as meetings. This is very important part of the evaluation of our system, but very difficult to execute as with too many users interacting at the same time the object detection system tends to become unstable and also so many users are difficult to control or guide while using the system. Also these situations results are very difficult to quantify, excluding the users reactions, so we are focusing on smaller, more controllable experiments.

The preliminary evaluation data we have show us that, even if the users need more time to perform a task, they feel the system is enjoyable and useful. We hope this advances will collaborate in bringing the tablespots environments outside the laboratories, evolving to a point in which can support everyday tasks in different fields. We want make possible to reduce the time the users need to learn to use and effectively use the tools so they can dedicate more time to creative tasks.

References

1. Pelaez, M.P., Choh, I.: Interactions with real and digital elements for collaborative document creation. In: Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work (2011)
2. Pelaez, M.P., Fujii, T., Nam, W., Yachiune, M., Choh, I.: Legible+: integrated system for remote collaboration through document creation. In: Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human, ICIS 2009 (2009)
3. Holman, D., Benko, H.: SketchSpace: designing interactive behaviors with passive materials. In: Proceedings of the 2011 Annual Conference Extended Abstracts on Human Factors in Computing Systems, CHI EA 2011 (2011)
4. Wellner, P.: The DigitalDesk calculator: tangible manipulation on a desktop display. In: Proceedings of the 4th Annual ACM Symposium on User Interface Software and Technology, UIST 1991 (1991)
5. Hartmann, M.R.M., Benko, H., Wilson, A.D.: Augmenting interactive tables with mice & keyboards. In: Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology, UIST 2009 (2009)
6. Benko, H., Morris, M.R., Brush, A.J.B., Wilson, A.D.: Insights on Interactive Tabletops: A Survey of Researchers and Developers. Microsoft Research Technical Report MSR-TR-2009-22 (March 2009)
7. Pauchet, A., Coldefy, F., Lefebvre, L., Louis Dit Picard, S., Perron, L., Bouguet, A., Collobert, M., Guerin, J., Corvaisier, D.: TableTops: Worthwhile Experiences of Collocated and Remote Collaboration. In: International Workshop on Horizontal Interactive Human-Computer Systems, TABLETOP 2007, pp. 27–34 (2007)
8. Jorda, S.: The reactable: tangible and tabletop music performance. In: Proceedings of the 28th of the International Conference Extended Abstracts on Human Factors in Computing Systems, CHI EA 2010. ACM, New York (2010)

9. Haller, M., Leithinger, D., Leitner, J., Seifried, T.: An augmented surface environment for storyboard presentations. In: Buhler, J. (ed.) ACM SIGGRAPH 2005 Posters, SIGGRAPH 2005. ACM, New York (2005)
10. Seifried, T., Jervis, M., Haller, M., Masoodian, M., Villar, N.: Integration of virtual and real document organization. In: Proceedings of the 2nd International Conference on Tangible and Embedded Interaction, Bonn, Germany, February 18 - 20 (2008)
11. Yoshida, T., Tsukadaira, M., Kimura, A., Shibata, F., Tamura, H.: Various tangible devices suitable for mixed reality interactions. In: Mixed and Augmented Reality, ISMAR (2010)
12. Nam, W., Fujii, T., Choh, I.: Legible Collaboration System Design. Hybrid Information Technology, ICHIT 2006 (2006)
13. Wobbrock, J.O., Morris, M.R., Wilson, A.D.: User-defined gestures for surface computing. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2009 (2009)
14. Weiss, M., Wagner, J., Jansen, Y., Jennings, R., Khoshabeh, R., Hollan, J.D., Borchers, J.: SLAP widgets: bridging the gap between virtual and physical controls on tabletops. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2009 (2009)

Using Kinect for 2D and 3D Pointing Tasks: Performance Evaluation

Alexandros Pino¹, Evangelos Tzemis^{2,3},
Nikolaos Ioannou², and Georgios Kouroupetroglou^{1,2}

¹ National and Kapodistrian University of Athens,
Accessibility Unit for Students with Disabilities, Athens, Greece

² National and Kapodistrian University of Athens,
Department of Informatics and Telecommunications, Athens, Greece
{pino, sdi0700116, koupe}@di.uoa.gr

³ University of Copenhagen, Department of Computer Science, Copenhagen, Denmark
khh107@alumni.ku.dk

Abstract. We present a study to comparatively evaluate the performance of computer-based 2D and 3D pointing tasks. In our experiments, based on the ISO 9241-9 standard methodology, a Microsoft Kinect device and a mouse were used by seven participants. For the 3D experiments we introduced a novel experiment layout, supplementing the ISO. We examine the pointing devices' conformance to Fitts' law and we measure a number of extra parameters that describe more accurately the cursor movement trajectories. Throughput, measured in bits per second is the most important performance measure. For the 2D tasks using Microsoft Kinect, Throughput is almost 39% lower than using the mouse, Target Re-Entry is 10 times up and Missed Clicks count is almost 50% higher. However, for the 3D tasks the mouse has a 9% lower Throughput than the Kinect, while Target Re-Entry and Missed Clicks are almost identical. Our results are also compared to older studies, and we finally show that the Kinect, operated by the user's hand and voice, is a suitable and effective input method for pointing and clicking, especially in 3D tasks.

Keywords: Fitts' law, 3D pointing, ISO 9241-9, Microsoft Kinect, Gesture User Interface.

1 Introduction

Nowadays, low-cost handheld devices introduced along with widespread game consoles can also be used as input devices in general purpose Personal Computers (PCs). Kinect [1] is a motion sensing input device by Microsoft for the Xbox 360 video game console and Windows PCs. Based around a webcam-style add-on peripheral, it enables users to control and interact with the Xbox 360 or the PC without the need to touch a game controller, through a natural user interface using gestures and spoken commands. The Kinect sensor is a horizontal bar connected to a small base with a motorized pivot and is designed to be positioned lengthwise above or below the video

display. The device features an RGB camera, depth sensor and multi-array microphone, running proprietary software, which provide full-body 3D motion capture, facial recognition and voice recognition capabilities. The Microsoft Kinect sensor, in opposition to most other accelerometer-based devices of this domain, uses the depth camera to recognize dynamic gestures, and this is the reason that the user does not need to use any kind of remote control, apart from his hands.

Although accelerometer-based recognition of dynamic gestures has been investigated in numerous studies (for examples see [2-3]), there is not an extensive research field of vision-based devices [4]. A promising software implementation that tracks the 3D position, orientation, and full articulation of a human hand from marker-less visual observations was developed by Oikonomidis, Kyriazis, and Argyros [5-6].

The abbreviations 2D and 3D in our case, where the Graphical User Interface (GUI) is displayed on a two-dimensional monitor in both cases, would more accurately be described by the terms two-directional and three-directional, rather than two-dimensional and three-dimensional.

The point-and-click metaphor usually referred to as pointing or tapping, constitutes a fundamental task for most 2D and 3D GUIs in order users to perform an object selection operation. Typing, resizing, dragging, scrolling, as well as other GUI operations require pointing. In order to develop better pointing techniques we need to understand the human pointing behavior and motor control. Fitts' Law [7] can be used to:

- Model the way users perform target selection
- Measure the user's performance,
- Compare the user's performance amongst various input devices or the change of performance over time.

Fitts' law has been applied to 3D pointing tasks [8] as well as to the design of gesture-based pointing interactions [9]. The most common evaluation measures related to Fitts' law are speed and accuracy, which are both incorporated in throughput.

2 Methodology

Fitts [7] proposed a model for the tradeoff between accuracy and speed in human motor movements. He proposed to quantify a movement task's difficulty using information theory by the metric of "bits". According to Fitts, the *Movement Time (MT)* needed to hit a target must be linearly related to the *Index of Difficulty (ID)* of the task:

$$MT = a + (b \times ID) , \quad (1)$$

where a and b are constants determined through linear regression,

$$ID = \log_2 \left(\frac{D}{W} + 1 \right) , \quad (2)$$

and D and W are the target's *Distance* and *Width* respectively.

Fitts proposed to quantify the human rate of information processing in aimed movements using “bits per second” as units. Fitts named the measure “index of performance”; today it is more commonly known as *Throughput* (TP , in bits/s). Although different methods of calculating *Throughput* exist in the literature, the preferred method is that proposed by Fitts in 1954 [7]. The calculation involves a direct division of means: dividing ID (bits) by the mean MT (seconds), computed over a block of trials:

$$TP = \frac{ID_e}{MT} \quad (4)$$

The subscript e in ID_e reflects a small but important adjustment, which Fitts endorsed in a follow-up paper [10]. The “adjustment for accuracy” involves first computing the *effective target Width* (W_e) as

$$W_e = 4,133 \times SD_x \quad (5)$$

where SD_x is the observed standard deviation in a participant's selection coordinates over repeated trials with a particular D - W condition. Computed in this manner, W_e includes the spatial variability, or accuracy, in responses. In essence, it captures what a participant actually did, rather than what he or she was asked to do. This adjustment necessitates a similar adjustment to ID , yielding an *effective Index of Difficulty*:

$$ID_e = \log_2 \left(\frac{D}{W_e} + 1 \right) \quad (6)$$

Calculated using the adjustment for accuracy, TP is a human performance measure that embeds both the speed and accuracy of responses. TP is most useful as a dependent variable in factorial experiments using pointing devices or pointing techniques as independent variables.

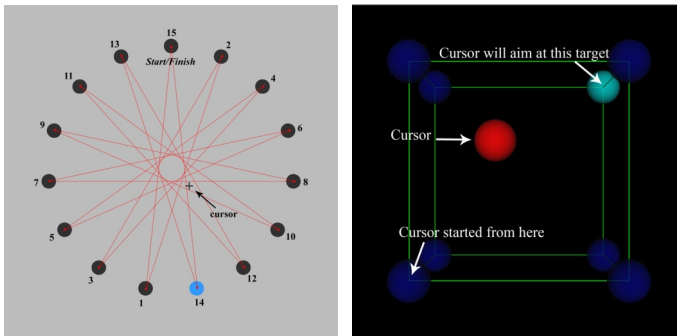
In order to evaluate the Kinect's conformance to Fitts' law as an input device, we used experimental software that we had previously designed and implemented [11], based on the ISO 9241-9 standard [12-13] that covers unidirectional and multidirectional user interaction. The ISO 9241-9 standard, describes a standardized procedure to evaluate the performance, comfort, and effort in using computer pointing devices; the use of the international standard grants the ability to better understand the experimental results, and to undertake comparisons between studies.

The GUIs of our 2D and 3D experiments are shown in Figure 1 left and right respectively. Users were asked to perform pointing tasks for 5 combinations of target distances and widths, hence 5 different ID s in increasing difficulty. For each ID they hit 15 targets in the 2D and 8 targets in the 3D experiment. Table 1 lists the ID s that were used for each of the 5 sessions. ID s were at a lower range in the 3D experiment because higher ID s resulted in non-displayable layouts either because the target distances were larger than the screen dimensions or because targets would be too small to see.

Table 1. Indexes of Difficulty (IDs) used for the 2D and the 3D experiments

Session	2D IDs	3D IDs
1 st	2.69	1.91
2 nd	3.19	2.35
3 rd	3.69	2.81
4 th	4.19	3.28
5 th	4.69	3.76

In the 2D case, the 15 circular targets are arranged in a circular layout (Figure 1, left). Initially the cursor is locked and unmovable on the center of the first target; the subject has to click in order to free the cursor and begin the experiment. Then the participant must move the cursor directly to the highlighted opposite target and click on it and so on clockwise. Each test block is complete when all 15 targets have been selected for 5 sessions giving a total of 75 trials per user. Circular targets and a cross-shaped cursor were used. The path the subject follows begins and ends at the top target. The lines in Figure 1 (left) indicate the ideal task path to alternating targets around the circle. Numbers indicate the succession of the targets to be hit. Software that captures the subjects' *MTs* and trajectory data, also graphically indicates which target the subject should proceed to (the lighter color target in the Figure). Figure 1 (left) illustrates the 4th ID experiment ($ID=4.19$)

**Fig. 1.** 2D (left) and 3D (right) pointing task screenshots

For the 3D experiment we have introduced a new layout, as the ISO 9241-9 did not contemplate for 3D interaction on the computer screen. We used 8 spherical targets placed at the vertices of a cube (Figure 1, right). Each task begins again with a click on the center of the first upper right target. Then the participant must move the cursor directly to the diagonally opposite target and click on it. After a successful trial the cursor teleports to another target that will become the beginning of the next route. The active target to be hit is highlighted every time. Each test block ends when all 8 equi-distance diagonal routes are successfully done (8 trials) and 5 sessions are run for different target radius and distance combinations (in total 5 different IDs) giving a total of 40 trials per user. The cursor was also a sphere (the red one). This way the user could perceive where the cursor is by its size in relation to the inner and outer

targets. The smaller the cursor the deeper it is inside the screen. When it is the same size as a target it means it is in the same z-axis level with it. Figure 1 (right) illustrates the experiment with $ID=3.28$.

We have to note that we disabled all windows cursor acceleration and accuracy enhancement options for the mouse, we used a 40×30 cm mouse pad, and the left mouse button for clicking (all users were right-handed). With the Kinect, for the 2D tests we acquired cursor movement coordinates by the depth camera data, after identifying the user's hand and taking as x and y the mean value of his hand's detected points. Movement in the z axis was not taken into account in 2D tests.

For the 3D experiments the difference was in that we used the mouse's scrolling wheel in order to move on the z axis, scrolling up to "go inwards" the screen and scrolling down in order to "come outwards". Regarding the Kinect, for 3D tests we used the z-axis data that were acquired through the Kinect's depth camera. We should also point out that in both experiments clicking was achieved using Kinect's microphones. The user had to produce a very short vowel phoneme in order for the click operation to be committed. He or she just had to say "Ah" for example in order to click.

Our analysis is based on the theory proposed by Fitts [7, 10] and MacKenzie et al. [14]. Specifically, we measured the following parameters (detailed definitions and formulas can be found in [11], [14], and [15]):

- *Throughput (TP)* in bits per second.
- Missed Clicks (MCL) scalar,
- Target Re-Entries (TRE) scalar,
- Task Axis Crossings (TAC) scalar,
- Movement Direction Changes (MDC) scalar,
- Orthogonal Direction Changes (ODC) scalar,
- Movement Variability (MV) in pixels,
- Movement Error (ME) in pixels,
- *Movement Offset (MO)* in pixels, and

Moreover, we have introduced a novel parameter **Distance Travelled (DT)**, defined as the distance travelled in pixels from the starting point to the successful click point inside the active target in each trial. It gives a sense of how close to the ideal path was the actual one. In a perfect trial where the cursor starts from the center of the starting target and the user clicks on the center of the active target, the *Distance Travelled* would be equal to the target's *Distance (D)*. Keates and al. [15]v also introduced a series of relative to the *DT* measures.

We developed the experimental application [16] as a Virtual Instrument using the LabVIEW (Laboratory Virtual Instrumentation Engineering Workbench) graphical programming environment by National Instruments [17]. We have tested Microsoft Kinect sensor as a gesture input device, and we developed appropriate software in LabVIEW for getting data in real time from its microphone array and the depth-camera.

The Microsoft Kinect was connected to a high-end PC using USB 2.0 communication. The computer was an Intel Core i7, 3.50 GHZ desktop with 8 GB of RAM

running MS-Windows 7 Professional and LabVIEW 2011. We used a 19'' TFT monitor with 1280×1024 resolution, and a 1600 dpi wireless mouse.

Seven (7) participants, volunteered for the study. Their age range was from 22 to 55 years. No one had any kind of disability and they had no experience with the Microsoft Kinect sensor.

Participants were instructed to try to hit the active target in each trial as fast and as closer to the center they could. In both tests users were instructed not to stop on erroneous clicks and an auditory feedback was given in that case. Visual feedback was also given when the cursor was in the target, and both auditory and visual feedback was given on successful clicks. In order for the users to achieve the best possible results they were instructed to wear a headset in order to reduce background noise and listen better to the audio feedback. Moreover, the experiments were taking place in a dark environment (no other source of light except for the TFT screen brightness), to avoid distraction caused by other objects in the environment and also for the screen to be more visible. For the mouse experiment users were sitting on a chair having a desk with the mouse and mouse pad in front of them. For the Kinect experiment they were standing up with their right hand outstretched in front at their chest level. In both experiments participants were situated 2m away from the screen. Each task was explained and demonstrated to participants and a warm up session at the medium *ID* was given for each device and each mode (2D, 3D).

3 Results

Measurements of the *Movement Time (MT)* as a function of the *Index of Difficulty (ID)* for all the participants in 2D and 3D experiments using the mouse and Microsoft Kinect sensor are presented in Figure 2.

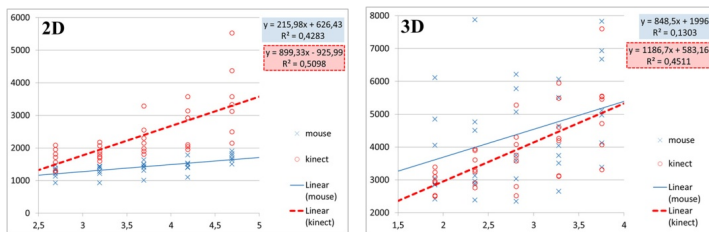


Fig. 2. Plots of the mean values of *Movement Time* for all trials (*MT* in milliseconds) as a function of the *Index of Difficulty (ID)* in bits) for all participants in 2D (left) and 3D (right) experiments using a Microsoft Kinect sensor (dashed lines) and the mouse (solid lines). Note: Value ranges in both axes differ between plots (axes are fitted to value ranged for better depiction).

The quantity *R*, called the *linear correlation coefficient*, measures the strength and the direction of the linear relationship between the two variables *ID* and *MT*. The value of the *correlation coefficient* is such that $-1 < R < +1$. Positive values indicate a relationship between the variables such that as values for *ID* increase, values for *MT*

also increase. Negative values indicate a relationship such that as values for *ID* increase, values for *MT* decrease. In our case *R* values were always positive. If there is no linear correlation or a weak linear correlation, *R* is close to 0. A value near zero means that there is a random, nonlinear relationship between the two variables. A perfect correlation of ± 1 occurs only when the data points all lie exactly on a straight line. A correlation greater than 0.8 is generally described as strong, whereas a correlation less than 0.5 is generally described as weak. In 2D experiments *R* was 0.65 for the mouse, and 0.71 for the Kinect. In 3D experiments *R* was 0.36 for the mouse and 0.67 for the Kinect.

The *coefficient of determination*, R^2 , is useful because it gives the proportion of the variance (fluctuation) of one variable that is predictable from the other variable. It is a measure that allows us to determine how certain one can be in making predictions from a certain model/graph, in our case Fitts' Law. The *coefficient of determination* is such that $0 < R^2 < 1$, and denotes the strength of the linear association between *ID* and *MT*. It represents the percent of the data that is the closest to the line of best fit. For example, the 2D Kinect line has $R^2=0.51$, which means that 51% of the total variation in *MT* can be explained by the linear relationship between *ID* and *MT* (as described by the regression equation). The other 49% of the total variation in *MT* remains unexplained. R^2 is a measure of how well the regression line represents the data, or how well our experiments' tasks comply with Fitts' Law. If the regression line passes exactly through every point on the scatter plot, Fitts' Law would be able to explain all of the variation. The further the line is away from the points, the less Fitts' Law is able to explain.

Table 2. Calculated parameters (means) of the cursor trajectory generated by the two input devices in 2D and 3D experiments.

		<i>TP</i>	<i>MCL</i>	<i>TRE</i>	<i>TAC</i>	<i>MDC</i>	<i>ODC</i>	<i>MV</i>	<i>ME</i>	<i>MO</i>	<i>DT</i>
2D	Mouse	3,45	0,30	0,07	2,62	31,50	1,24	11,39	12,95	3,73	569
	Kinect	2,10	0,28	0,69	4,85	18,82	5,28	16,97	15,80	3,41	905
3D	Mouse	0,96	0,53	0,18	-	41,10	12,37	78,30	109,10	-	1285
	Kinect	1,06	0,28	0,66	-	28,08	25,07	80,83	106,53	-	1563

Table 2 presents the results of the statistical analysis of all data from all users 2. We note that in 3D experiment *Task Axis Crossing* has no meaning because even if there is an axis connecting the two sphere targets, there is a too low possibility to actually cross it, and also that *Movement Error* is identical to *Movement Offset*.

Figure 3 illustrates typical sessions observed for the mouse (left) and the Kinect (right). It is obvious that for 2D tasks the mouse shows smoother and more accurate behavior. Kinect introduces tremor that typically appears in most people's hands; even when trembling is unnoticeable by eye, the plots of the pointing tasks reveal it.

For the specific sessions illustrated in Figure 3 we also give the mean values of the most important measures for comparison:

- Mouse: $MT = 1829$ ms, $TP = 3.27$ bits/s, $DT = 783.75$ pixels
- Kinect: $MT = 2148$ ms, $TP = 2.60$ bits/s, $DT = 865.61$ pixels

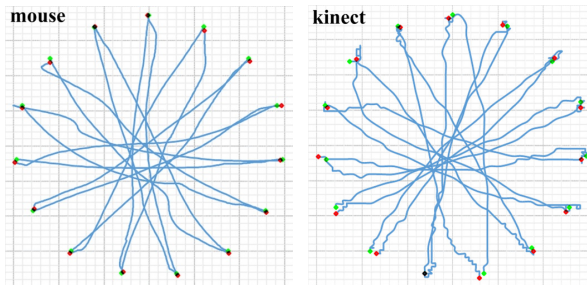


Fig. 3. Plots of the most difficult 2d session with $ID=4.69$: mouse (left), and Kinect (right)

We will briefly compare the results between this experiment and our previous one, which comprised the mouse and the Wiimote [16]. For the mouse, the basic difference is that in our previous experiments the mouse was enhanced by enabling the relevant windows settings; in Windows 7 the setting is “Enhance pointer precision”, and it accelerates the cursor when the user moves it fast (meaning he is “travelling” towards a target that is still far away), while it decelerates the cursor when the user is moving it slower (meaning he has reached close to the target and tries to click on the right spot). This way, in our previous experiments, users had some virtual aid when they were trying to hit the targets with the mouse. However, in order for the comparison between devices to be fair, we decided in the current experiment to use the mouse without any enhancements. We also adjusted the mouse speed in order to be comparable with the other device’s speed. Making ergonomic measurements we set up the experiment so that the whole range of the mouse’s movement (depending on the users’ arm and hand movement capabilities on the table) corresponds to the whole range of the Kinect movement (depending on the users’ arm and hand movement in the air) and both ranges have the same effect on the cursor’s movement on the screen.

In the previous experiment the mouse had a *Throughput* of 5.05 for the 2D and 1.71 for the 3D experiment, which were much better measurements than the 3.45 and 0.96 we respectively got now. The previous accuracy enhancement settings that were disabled now fully justify this difference.

Nevertheless, what is of higher interest and importance is the comparison of the Wiimote and Kinect performances: In our past experiments Wiimote gave a *Throughput* of 2.97, and 0.75 in 2D, and 3D tasks respectively. In the current experiment Kinect gave 2.10, and 1.06. This means that the Wiimote is better in 2D tasks and Kinect is better in 3D pointing.

4 Conclusion

Based on the ISO 9241-9 standard, we have presented an experimental study for the comparatively evaluation of computer-based 2D and 3D performance pointing tasks. A Microsoft Kinect device and a mouse were used by seven participants. For the 3D experiments we introduced a novel experiment layout, supplementing the ISO as well as the novel evaluation parameter **Distance Travelled (DT)**.

We conclude that for the 2D tasks using the Microsoft Kinect sensor, *Throughput*, is 39% lower than using the mouse, *Missed Click* count almost the same. However, for the 3D tasks using the Microsoft Kinect sensor, *Throughput* is 9.7% higher than using the mouse, while *Missed Clicks* are considerably lower.

Furthermore, Figure 2 shows that the fitting line *coefficient of determination (R^2)*, which reflects the reliability of the linear relationship between *MT* and *ID* values and, therefore, the compliance to Fitts' law, is slightly higher for the Kinect device than the mouse for the 2D experiment. When it comes to the 3D experiment, we can definitely observe that the compliance with Fitts' law is now much higher for the Kinect than the mouse.

Finally, we round off that the Microsoft Kinect device was proven to be a slower and harder to use input device for the 2D pointing task compared to the mouse. However 3D tests show that the condition is reversed as far as speed and accuracy of the Kinect are concerned. Kinect works better than the mouse in 3D. However, we can argue that both the mouse and the Kinect had too low *TP* as 3D pointing devices, which is partly justified by the fact that all users had no previous experience of any kind of 3D interaction.

Future work will include the involvement of more users in the experiments, also disabled users, research on how performance changes over time (i.e., familiarization with the Kinect and performance improvement), and introduction of new trajectory measures for 3D tasks in spherical coordinates.

Acknowledgments. The work described in this paper has been funded by the Special Account for Research Grants of the National and Kapodistrian University of Athens.

References

1. Microsoft: Kinect - Xbox.com. Xbox 360+Kinect homepage, <http://www.xbox.com/en-US/kinect>
2. Kela, J., Korpipää, P., Mäntyjärvi, J., Kallio, S., Savino, G., Jozzo, L., Marca, D.: Accelerometer-Based Gesture Control for a Design Environment. *Pers. Ubiquit. Comput.* 10(5), 285–299 (2006)
3. Kratz, S., Rohs, M.: The \$3 recognizer: Simple 3D Gesture Recognition on Mobile Devices. In: *IUI 2010, 15th International Conference on Intelligent User Interfaces*, pp. 419–420. ACM Press, New York (2010)
4. Marvel, J.A., Franaszek, M., Wilson, J., Hong, T.H.: Performance Evaluation of Consumer-Grade 3D Sensors for Static 6DOF Pose Estimation Systems. In: Tescher, A.G. (ed.) *Applications of Digital Image Processing XXXV, SPIE Optical Engineering and Applications Conference*, vol. 8499, article 849905. SPIE, Bellingham (2012)

5. Oikonomidis, I., Kyriazis, N., Argyros, A.A.: Efficient Model-Based 3D Tracking of Hand Articulations Using Kinect. In: Hoey, J., McKenna, S., Trucco, E. (eds.) 22nd British Machine Vision Conference, pp. 101.1–101.11. BMVA Press, Manchester (2011)
6. Oikonomidis, I., Kyriazis, N., Argyros, A.A.: Tracking the Articulated Motion of Two Strongly Interacting Hands. In: CVPR 2012, IEEE Conference on Computer Vision and Pattern Recognition, pp. 1862–1869. IEEE (2012)
7. Fitts, P.M.: The information capacity of the human motor system in controlling the amplitude of movement. *J. Exp. Psychol.* 47(6), 381–391 (1954)
8. Murata, A., Iwase, H.: Extending Fitts' law to a three-dimensional pointing task. *Hum. Mov. Sci.* 20, 791–805 (2001)
9. Foehrenbach, S., König, W.A., Gerken, J., Reiterer, H.: Tactile Feedback Enhanced Hand Gesture Interaction at Large, High-resolution Displays. *J. Visual Lang. Comput.* 20(5), 341–351 (2009)
10. Fitts, P.M., Peterson, J.R.: Information capacity of discrete motor responses. *J. Exp. Psychol.* 67(2), 103–112 (1964)
11. Pino, A., Kalogeros, E., Salemis, I., Kouroupetroglou, G.: Brain Computer Interface Cursor Measures for Motion-impaired and Able-bodied Users. In: Stephanidis, C. (ed.) *HCI International 2003. Universal Access in HCI: Inclusive Design in the Information Society*, vol. 4, pp. 1462–1466. Lawrence Erlbaum Associates, Mahwah (2003)
12. ISO 9241-9:2000: Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs) - Part 9: Requirements for Non-keyboard Input Devices. ISO Standard (2000)
13. Soukoreff, W.R., MacKenzie, I.S.: Towards a standard for pointing device evaluation, perspectives on 27 years of Fitts' law research in HCI. *Int. J. Hum.-Comput. St.* 61(6), 751–789 (2004)
14. MacKenzie, I.S., Kauppinen, T., Silfverberg, M.: Accuracy measures for evaluating computer pointing devices. In: *CHI 2001, SIGCHI Conference on Human Factors in Computing Systems*, pp. 9–16. ACM Press, New York (2001)
15. Keates, S., Hwang, F., Langdon, P., Clarkson, J.P.: Cursor measures for motion-impaired computer users. In: *ASSETS 2002, the 5th International ACM Conference on Assistive Technologies*, pp. 135–142. ACM Press, New York (2002)
16. Kouroupetroglou, G., Pino, A., Balmpakakis, A., Chalastanis, D., Golematis, V., Ioannou, N., Koutsoumpas, I.: Using Wiimote for 2D and 3D Pointing Tasks: Gesture Performance Evaluation. In: Efthimiou, E., Kouroupetroglou, G., Fotinea, S.-E. (eds.) *GW 2011. LNCS (LNAI)*, vol. 7206, pp. 13–23. Springer, Heidelberg (2012)
17. National Instruments: NI LabVIEW – Improving the Productivity of Engineers and Scientists. LabVIEW System Design Software homepage, <http://www.ni.com/labview/>

Conditions of Applications, Situations and Functions Applicable to Gesture Interface

Taebeum Ryu¹, Jaehong Lee², Myung Hwan Yun², and Ji Hyouon Lim³

¹ Department of Industrial and Management Engineering,
Hanbat National University, Daejeon, 305-719
tbryu75@gmail.com

² Department of Industrial Engineering,
Seoul National University, Seoul, 151-741
{eyed04, mhy}@snu.ac.kr

³ Department of Industrial Engineering, Hongik University, Seoul, 121-791
limhj@hongik.ac.kr

Abstract. Although there were many studies related to developing new gesture-based devices and gesture interfaces, it was little known which applications, situations and functions are applicable to gesture interface. This study developed a hierarchy of conditions of applications (devices), situations and functions which are applicable to gesture interface. This study searched about 120 papers relevant to designing and applying gesture interfaces and vocabulary to find the gesture applicable conditions of applications, situations and functions. The conditions which were extracted from 16 closely-related papers were rearranged, and a hierarchy of them was developed to evaluate the applicability of applications, situations and functions to gesture interface. This study summarized 10, 10 and 6 conditions of applications, situations and functions, respectively. In addition, the gesture applicable condition hierarchy of applications, situation and functions were developed based on the semantic similarity, ordering and serial or parallel relationship among them. This study collected gesture applicable conditions of application, situation and functions, and a hierarchy of them was developed to evaluate the applicability of the gesture interface.

Keywords: Gesture interface, Applicability, Gesture application, Situation, Functions.

1 Introduction

As different devices for advanced technologies such as smart-home systems, robots, and large screen displays appear in the market, the demands for more applicable interfaces have increased for such devices. Moreover, recent technologies have allowed the consumers to have more intuitive interactions with the devices, and the gesture interface is one of those recent technologies that have been introduced [21].

There has been an active movement in the recent literature on how to incorporate the gestures with various interface technologies. The earlier research studies have

focused on developing new devices and applications that incorporate such gesture interfaces [2] while recent studies are more focusing on the architecture of gesture vocabulary and the evaluation of gesture applicability conditions [9].

However, the current literature still lacks the research on understanding which applications, situations and functions (or commands) to incorporate such gesture interfaces. The current research in the field of ergonomics has conducted studies on developing more intuitive gestures and evaluating the applicability of such gestures. However, in order to increase the utility of the gesture interfaces, more studies on determining which applications, situations and functions to apply the gesture interfaces need to be first investigated before developing and evaluating the gestures themselves.

This study has gathered and systemized the conditions of the device, situation and function to appropriately apply the gesture. From studying the existing studies related the gesture interface, we have collected the information on the conditions of the device, situation and function for the gesture applicability and have developed a system for them. The conditions derived from our study may be utilized as categories to evaluate the applicability of the gesture for a device, situation and function prior to designing the gestures. However, detailed outline and guidelines are necessary in order to utilize the conditions in practice.

2 Method

In order to develop a guideline for evaluating the gesture applicability, this study has collected the information from the existing studies on the design, evaluation and application of the gesture interface and has derived the categories for the conditions of the device, situation and function to apply the gesture interface. About 120 studies have been considered, but only about 16 of them were closely related to the concern of our study. Wachs et al. [21], which concerns the necessary conditions of the generalization of the interface, has also mentioned the lack of such studies in the literature

This study has utilized the categories derived from the existing studies to organize and systemize the conditions of the device, situation and function to apply the gesture interface. The collected conditions were categorized according to the similar meanings and the level of the applicability.

3 Application Conditions

3.1 Hierarchy of Application (Device) Conditions

The conditions of a device for applying the gesture interface are categorized into 10 different criteria: price and cost, communication, behavioral pattern, entertainment, security, spatial manipulation, remote control, urgency, sterility, and support for elderly and disabled.

Figure 1 shows the hierarchy of device conditions. Except the price and cost, the rest of the conditions are categorized into three characteristics of a gesture: naturalness, expressiveness, and contact-free. Naturalness concerns communication and behavioral pattern, while expressiveness concerns entertainment, security, and spatial manipulation. Contact-free concerns remote control, urgency, sterility, and support for elderly and disabled. The further discussion of the device conditions are found in Ch. 3.2 through Ch. 3.5, while Ch. 3.6 explains the usage of the device conditions for evaluating the applicability of the gesture interface.

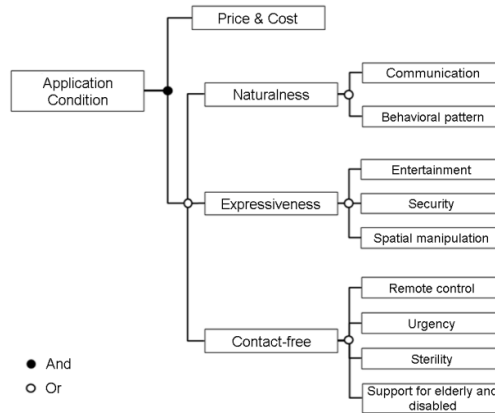


Fig. 1. Hierarchy of application (device) conditions applicable to gesture interface

3.2 Price and Cost

The production cost and the price of the devices should fall within the reasonable range in regard to the gestures that are applied. In other words, the production cost with the gestures applied should not be higher than those without, and the efficiency of the final devices should be worth more than the cost of the production [9].

3.3 Naturalness (Natural Interaction)

The devices require natural communication with the gestures; gestures refer to the physical movement acted during communication with oneself or others [5]. Gestures are naturally accompanied by verbal communication during interactions among men [1]. Therefore, the devices, to which the gestures are applied, should also find such gestures useful for the communication with the users. For example, robots are often aided by the gestures of the human users to successfully accomplish the tasks. Vacuum cleaning robots such as Roomba and robot pets such as AIBO are further developed with control interfaces for allowing natural gestures of the human users.

Moreover, certain applications detect and recognize different conditions of the users such as drowsiness and excitement through their natural depiction of gestures. The emotions and physiological conditions are often depicted through the users' head

movements and face expressions [11]. For example, drowsiness is often recognized by the user's head movement while surprise or anger could be recognized by the face expression. Such gestures are considered passive and can be utilized to recognize false witness, nervousness, drowsiness, distress, and related conditions and emotions.

3.4 Expressiveness

The devices that are in need of the gesture interface should allow diverse expressions of gestures and have a certain purpose of entertainment. As Mitra and Acharya [11] have defined the gesture as a meaningful and expressive body movement, the gesture itself has a purpose of expression. Gestures that are focused on expressions convey different human emotions [11], while offering different pleasures of expression (e.g., dance and performing arts), and are also used for showing individuality of the performers [8]. Therefore, the devices should be also in desire of expressive gestures.

The most common cases of such devices are video game consoles. Gaming devices like Nintendo Wii and Xbox Kinetic have gained the popularity for their diverse range of expressive gestures. Moreover, the individuality conveyed through the gestures is also utilized in security systems such as door locks and alarm systems [8].

In addition, another meaningful information of the gestures is spatial information, and the devices should also be benefited from such information. Gestures provide spatial information such as location and direction and can be most utilized for controlling an object in space. Particularly, the gestures may be most effective than other interfaces in three-dimensional space [16], and large and high screen context which deals direct manipulation are most benefited from the gesture interface [3].

Devices that may be benefited by the gesture interface are visual architecture applications, 3D virtual reality systems, Smart Design Studio [8], and other related large-screen systems.

3.5 Contact-free

Since the gestures are often made without physical contact of the devices, contact-free is a basic characteristic of the devices for the gesture interface. Although certain applications of touch-screens are utilizing the gestures that are in contact of the device, the demand of contact-free gestures would increase in other ubiquitous environments. The contact-free gestures are beneficial in terms of remote control, urgency, mobility, sterility, and support for elderly and disabled [22].

The examples of the applications that utilize the contact-free characteristic of the gestures are smart TVs linked with smart-home systems, emergency systems, medical devices, and disability aids [22]. TVs often require remote control, and there have been studies on developing such applications centralized by smart-home systems [9, 14]. Moreover, emergency system has been developed for higher effectiveness through the gestures that do not require physical contact of the controllers for various emergency situations [17]. Medical devices require cautious usage due to the possible infections, and certain devices such as FACE MOUSE [13] have been introduced to utilize the face expressions into the gesture interface. Lastly, wheelchairs have been

developed to aid the disabled people to manipulate more easily without much of the muscle strengths [7].

3.6 Usage of Conditions in Evaluating Applications

The devices must satisfy the condition of the production cost and product price, as well as at least one of the rests of the conditions. Therefore, as Figure 1 shows, the evaluation logic has been established with the AND condition between the production cost and the rest of the conditions (naturalness, expressiveness and contact-free), which have the relationship of the OR condition. Likewise, the lowest level of the conditions of naturalness, expressiveness and contact-free must also go through the OR condition to have at least one of them satisfied.

4 Situational Conditions

4.1 Hierarchy of Situational Conditions

A total of 9 conditions have been collected from the existing studies for the situations to apply the gesture interface: physical availability, attention resource, space, posture, recognition error reduction, noise and interference, social acceptance, visibility, and minimal speech.

These situational conditions have been organized into two large categories of allowance and necessity, as shown in Figure 2. The further explanations are discussed in Ch. 4.2 and 4.3, and the usage of the conditions is discussed in Ch. 4.4.

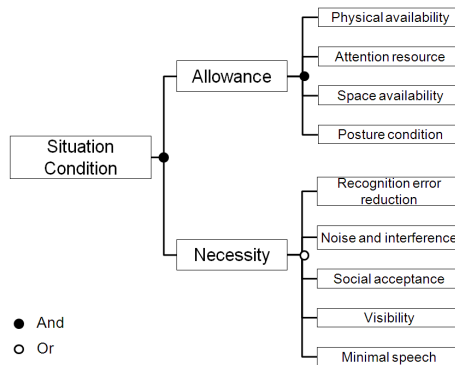


Fig. 2. Hierarchy of situation conditions applicable to gesture interface

4.2 Allowance Condition

The first situational conditions for the gesture interface to be applied are the allowance of the situations. The allowance conditions comprise of physical availability, attention resource, space, and posture. First, physical availability refers to

the availability of the body parts for gestures such as hands and heads; if such body parts are already occupied to manipulate the device, this condition is not met. For example, if the user is carrying a device with both of his hands, then physical availability is quite low for the hands.

Attention resource refers to the availability of the resources for the user to pay attention to his gestures; if the user is preoccupied with tasks to process the spatial information or is going through manual tasks that require the user's immediate responses, this condition is not met [23]. Therefore, attention resource is a necessary condition among others to satisfy the situational conditions.

Moreover, the situation must allow sufficient space for the users to make gestures and postures.

4.3 Necessity Condition

The second situational conditions for the gesture interface are the situations where the interface is necessary. The necessity conditions consist of recognition error reduction, noise and interference, social acceptance, visibility, and minimal speech. Gestures are advantageous when there exists large errors in multimodal interactions while recognizing voice commands; the gestures provide another modality to the system to reduce the recognition error and thus are necessary in such situations [16]. Gestures are also necessary in situations where there exists noise or an obstacle that interferes the voice commands [19]. Moreover, there may be situations when the social acceptance is low for voice commands to be at a high volume [18]. Lastly, situations where visual information is lacking, such as in dark places, and where it requires minimal speech are both in need of the gesture interface [15].

4.4 Usage of Conditions in Evaluating Situations

The situations must satisfy both the allowance conditions and the necessity conditions. Therefore, these two groups of conditions are evaluated under the AND condition as shown in Figure 2. The gesture interface may only be applied when all of the allowance conditions are met; thus, these conditions are evaluated under the AND condition as well. Meanwhile, only one of the necessity conditions satisfy the situation to have the gesture interface applied, and they are evaluated under the OR condition.

5 Functional Conditions

5.1 Hierarchy of Functional Conditions

In the case of functional conditions, there have been a total of 7 conditions collected from the existing studies: simple and basic, repletion in short time, short-cut, immediateness, spatial information, spatial metaphor, and symbolic meaning.

These conditions are grouped into three categories of allowance, necessity and easy expression as shown in Figure 3. The allowance conditions are simple and basic and repetition in short time. The necessity conditions are short-cut and immediateness, and the easy expression conditions are spatial information, spatial metaphor and symbolic meaning. The further explanations of the conditions discussed in Ch. 5.2 through 5.4, and the usage of the conditions in evaluation is discussed in Ch. 5.5.

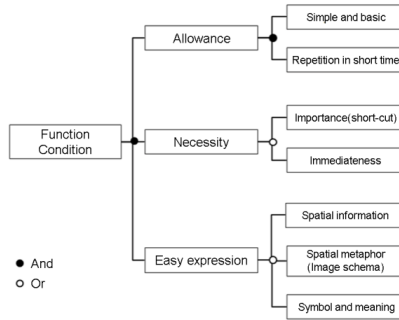


Fig. 3. Hierarchy of function conditions applicable to gesture interface

5.2 Allowance Condition

Allowance condition refers to a function that is simple and basic and is supplied in short time in order to apply the gesture interface. Simple and basic functions refer to single-unit tasks followed by immediate responses and results [8]. For example, tasks such as turning on/off a TV and increasing/decreasing audio volume are basic tasks that do not require further process of functions. Functions that request continuous gestures for more than one process of tasks are often complicated and have trouble recognizing the gestures. The users also find the gestures difficult to make in such cases.

Repletion in short time means lower frequency of the tasks. Gestures are highly involved with the user's muscle movements and thus may be tiresome if repeated for several times within a short period of time [24]. Therefore, repletion in short time is necessary for the gesture interface so that the users do not become fatigued [12]. For example, frequent change of channels may easily tire the users.

5.3 Necessity Condition

The functions to which the gesture interface is applied must be important and must require immediate reactions. Such functions should be used more frequently than others and should benefit from having short-cut features [20]. For example, remote-controlling TVs occurs frequently and is an important task, and thus a gesture interface is necessary.

Moreover, such functions must be in need and urgency of immediate tasks [8]. For example, the mute function is an immediate task of reducing the volume of a TV and would benefit from having a gesture interface.

5.4 Easy Expression

Lastly, the functions should be easy to replace with gestures to benefit from a gesture interface. Such functions deal with spatial information, spatial metaphor and symbolic meanings. The functions that are closely related with spatial information and movement information are easy to incorporate with a gesture interface [16]. For example, selection of a target, change of a location, size manipulation, and direction change are easy to apply the gesture interface.

Functions that could be related to abstract concept of the space are also easy to replace with gestures. For example, studies like Hurtienne et al. [6] have shown that concepts such as good/bad, close/far, center/circumference, and forward/backward are often explained with spatial metaphors and may be easily replaced by gestures.

Moreover, functions with symbolic meanings are easy to apply the gesture interface. For example, functions related to letter editing, and numbers or symbols could be easily replaced by gestures [16, 8].

5.5 Usage of Conditions in Evaluating Functions

The functional conditions must be satisfied to apply the gesture interface in terms of allowance, necessity and easy expression. Therefore, all three groups of conditions are evaluated under the AND condition as shown in Figure 3. Also, the further conditions of allowance must be all satisfied and thus are evaluated under the AND condition. On the other hand, at least one of the necessity conditions and of easy expression conditions should be satisfied and thus are evaluated under the OR condition.

6 Discussion and Conclusions

This study has proposed a development of systematic evaluation of the device (application), situational and functional conditions prior to incorporating a gesture interface. The evaluation of such conditions may benefit to provide easier tools to design the gesture interface, but research studies on developing such evaluation methods are still lacking in the literature. The conditions introduced in this study are collected from the context of the existing studies and thus are reliable resources to propose a systematic approach of evaluating the conditions to apply the gesture interface.

However, the evaluation method of the conditions discussed in this study still suggests further improvement. First, the conditions collected through this study do not comprehend all the possible conditions to evaluate the applicability of the gesture interface; thus, further investigation on collecting data is necessary. Second, further

discussion and validation are necessary to confirm that the collected conditions are the most appropriate to evaluate the applicability of the gesture interface. Moreover, further research is necessary to validate the system proposed in this study to categorize the conditions.

This study introduces the following process to utilize the device, situational and functional conditions for evaluating the applicability of the gesture interface. First, a complete mapping of the conditions must be developed for objective and comprehensive evaluation. The conditions are ambiguous and abstract to be appropriately evaluated. For example, frequency of the functions may be utilized as evaluation methods to measure the weight of functional necessity. Second, further categorization is necessary for objective evaluation of the conditions. For example, in the case of evaluating the necessity of the functions, detailed guidelines must be present in order to quantify the frequency of a function usage.

The further direction of the study focuses on improving the system of evaluating the applicability of the gesture interface. Based on the developed system, further methods of evaluating the device, situation and function will be developed. Moreover, validation of the system will be conducted with practical devices, situations and functions. Further development of the methods to evaluate the conditions themselves will also be considered.

Acknowledgements. This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2012-0003457).

References

1. Bhuiyan, M., Picking, R.: Gesture-controlled user interfaces, what have we done and what's next? In: 5th Collaborative Research Symposium on Security, E-Learning, Internet and Networking, pp. 59–60 (2009)
2. Bhuiyan, M., Picking, R.A.: Gesture controlled user interface for inclusive design and evaluative study of its usability. *Journal of Software Engineering and Applications* 4, 513–521 (2011)
3. Blatt, L., Schell, A.: Gesture Set Economics for Text and Spreadsheet Editors. In: *Proceedings of the Human Factors and Ergonomics Society 34th Annual Meeting*, pp. 410–414 (2011)
4. Guo, C., Sharlin, E.: Exploring the use of tangible user interfaces for human-robot interaction: a comparative study. In: *CHI 2008: Proceeding of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems*, pp. 121–130 (2011)
5. Hummels, C., Stappers, P.J.: Meaningful gestures for human computer interaction: beyond hand postures. In: *Proceeding of Third IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 591–596 (1998)
6. Hurtienne, J., Stöbel, C., Sturm, C.: Physical gestures for abstract concepts: Inclusive design with primary metaphors. *Interacting with Computers* 22(6), 475–484 (2010)
7. Jia, P., Hu, H., Lu, T., Yuan, K.: Head gesture recognition for hands-free control of an intelligent wheelchair. *Industrial Robot: An International Journal* 34(1), 60–68 (2007)

8. Kela, J., Korpipää, P., Mäntyjärvi, J., Kallio, S., Savino, G., Jozzo, L., Marca, D.: Accelerometer-based gesture control for a design environment. *Personal and Ubiquitous Computing* 10(5), 285–299 (2006)
9. Kühnel, C., Westermann, T., Hemmert, F.: I'm home: Defining and evaluating a gesture set for smart-home control. *International Journal of Human-Computer Studies* 69(11), 693–704 (2011)
10. Li, J.: Communication of Emotion in Social Robots through Simple Head and Arm Movements. *International Journal of Social Robotics* 3, 125–142 (2010)
11. 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)
12. Nielsen, M., Störring, M., Moeslund, T.B., Granum, E.: A procedure for developing intuitive and ergonomic gesture interfaces for man-machine interaction. In: *Proceedings of the 5th International Gesture Workshop*, pp. 1–12 (2003)
13. Nishikawa, A., Hosoi, T., Koara, K., Negoro, D., Hikita, A., Asano, S., Kakutani, H., Miyazaki, F., Sekimoto, M., Yasui, M., Miyake, Y., Takiguchi, S., Monden, M.: FACE MOUSE: A novel human-machine interface for controlling the position of a laparoscope. *IEEE Transactions on Robotics and Automation* 19(5), 825–841 (2003)
14. Neßelrath, R., Lu, C., Schulz, C.H., Frey, J., Alexandersson, J.: A gesture based system for context-sensitive interaction with smart homes. In: *Wichert, R., Eberhardt, B. (eds.) Advanced Technologies and Societal Change*, pp. 209–219. Springer, Berlin (2011)
15. Oviatt, S., DeAngeli, A., Kuhn, K.: Integration and synchronization of input modes during multimodal human-computer interaction. *Referring Phenomena in a Multimedia Context and their Computational Treatment*, 1–13 (1997)
16. Oviatt, S.: Ten Myths of Multimodal Interaction. *Communications of the ACM* 42(11), 74–81 (1999)
17. Rauschert, I., Agrawal, P., Sharma, R., Fuhrmann, S., Brewer, I., MacEachren, A.M.: Designing a human-centered, multimodal GIS interface to support emergency management. In: *Proceedings of the 10th ACM International Symposium on Advances in Geographic Information Systems*, pp. 119–124 (2002)
18. Rico, J.: Usable gestures for mobile interfaces: evaluating social acceptability. In: *Proceedings of the 28th International Conference on Human Factors in Computing System*, pp. 887–896 (2010)
19. Ronkainen, S., Koskinen, E., Liu, Y., Korhonen, P.: Environment Analysis as a Basis for Designing Multimodal and Multidevice User Interfaces. *Human-Computer Interaction* 25(2), 148–193 (2010)
20. Rhyne, J.: Dialogue Management for Gestural Interfaces. *Computer Graphics* 21(2), 137–142 (1987)
21. Shan, C.: Gesture Control for Consumer Electronics. In: *Multimedia Interaction and Intelligent User Interfaces*, pp. 107–128 (2010)
22. Wachs, J., Kolsch, M., Stern, H.: Vision-based hand-gesture applications. *Communications of the ACM* 54(2), 60–71 (2011)
23. Wickens, C.D., Hollands, J.G.: *Engineering psychology and human performance*, 3rd edn. Prentice Hall (1999)
24. Wilson, A., Oliver, N.: GWindows: Towards Robust Perception-Based UI. In: *First IEEE Workshop on Computer Vision and Pattern Recognition for Human Computer Interaction* (2003)
25. Young, J., Sung, J., Voids, A., Sharlin, E.: Evaluating human-robot interaction. *International Journal of Social Robotics* 3, 53–67 (2011)

Communication Analysis of Remote Collaboration System with Arm Scaling Function

Nobuchika Sakata, Tomoyuki Kobayashi, and Shogo Nishida

Division of Systems Science and Applied Informatics Graduate School of Engineering Science,
Osaka University 1-3, Machikaneyama-cho, Toyonaka-city, Osaka, 560-8531 Japan
{sakata, kobayashi, nishida}@sys.es.osaka-u.ac.jp

Abstract. This research focuses on the remote collaboration in which a local worker works with real objects by a remote instructor. In this research area, there are some systems which consist of the ProCam system consisting of a camera and a projector at the work environment and the tabletop system consisting of a display, a depth sensor and a camera at remote instructor environment. As the function enhancement, the system using the scaling method of the embodiment exists. The system makes it possible for instructor to instruct smoothly even to small objects and has an effect on task completion time in the user study of putting smaller block clusters than the size of fingers. We first analyzed the movie of previous experiment again, and then find out the problems the previous work could not solve, and proposed their solution.

Keywords: Remote collaboration, Scaling Method and Video Analysis.

1 Introduction

Work conducted by a local worker under the instructions of a remote instructor is called remote collaboration. Using a telecommunication terminal, the remote instructor and the local worker transmit and receive sounds and videos to accomplish their work since they cannot share voices and views directly. On the other hand, a worker and an instructor sometimes communicate regarding objects and places in real work spaces in local collaborative works [1][2][3]. To conduct such communication smoothly, a support system sends the remote instructor's instructions including the place of the work to the local worker.

Especially, some studies focus on the situation in which a remote instructor provides an instruction to a local worker with real objects, for example, repairing machinery. In these studies, a tabletop display is adopted to capture the gesture of the instructor and a projector is adopted to project the gesture image to the real-world directly. With these devices, it becomes easy that a local worker realizes an instruction intuitively with watching the projected image of instruction gesture on the work environment.

Uemura[4] proposed and developed the remote collaborative work system with the scaling method of the body image as an instruction image. Then, it studied and confirmed the efficiency of his proposed method in terms of the task completion time

and the result of questionnaire by conducting the user study of putting block clusters. However, resulting from the re-discussion of previous work, we found some issues about the rate of system utilization and questionnaires. Therefore, we first analyzed the movie of experiment of previous work particularly and cleared the issues of previous work. In parallel, we picked up the scenes when a worker did not work smoothly even when the scaling system was used and picked up the difference between with and without the scaling method. We discussed the problems we found in the movies and proposed the solutions for them. After that, we implemented the function that instructor's device could save the image of worker environment as instructor liked and could display it blended with the current condition of worker environment. Lastly, we conducted the user study to examine the effectiveness of the proposal method.

2 Related Work

Some research studies the support of the instruction to the local worker by the remote instructor as a remote collaboration. Some of these research focus on the remote collaboration with real-world objects. The tele-operated laser pointer is adopted in some research as a pointing tool for remote collaboration[5][6][7]. Cterm[5] and Gesture-Laser[6] are device placed in a work space, and WACL[7] is a wearable device. Each of these is compact size and consists of a camera, a microphone, a speaker and a laser pointer which are remotely controlled. The instructor can pan and tilt the laser pointer on the camera to point at real-world objects. GestureMan[8] is a system equipped with not only a tele-operated laser pointer but also a robot head and a robot arm. The robot head and the robot arm trace the motion of the remote instructor.

Kondo[9] develops view sharing system between an instructor and a worker for remote collaboration. This system is constructed from the video-see-through Head Mounted Displays(HMD) and motion trackers. The system allows two users in remote places to share their first-person views each other. To achieve the instruction considering embodiment in the remote collaboration, some research display the image or the shadow of the instructor on the work environment[10][11][12][13].

These research show the remote communication becomes smooth by considering embodiment and transmitting the awareness information or gestures. Therefore, the instruction via instruction images is effective for the remote collaboration with real world objects. Moreover, considering embodiment and transmitting gesture or awareness information is important in the instruction with real-world objects. However, above systems focus on the system placed on the work environment. There has been some researches which propose the system and the remote interaction for the instructor.

3 Previous Work

As the base of this research, the remote collaborative work system has been developed by our colleague. Uemura[4] proposed with the scaling method of the body image as an instruction image. In this chapter, we introduce the system and the user study of its work.

3.1 System Overview and Method of Instruction

System of previous work has two interfaces. One is the instructor interface for remote place as shown in left side of Figure 1. The other is the worker interface using by local worker as shown in right side of Figure 1. The instructor interface consists of tabletop display (byd:sign, d:3232GJC3 32Inch) as an output device, RGB camera and depth sensor (Microsoft, Kinect) as an input device. The worker interface consists of a micro projector (MITSUBISHI, LVP-XD95) as an output device and a camera(Point Grey Inc., Firefly FMVU-03MTC-CS) as an input device. Next, we state the instruction sequence using the hardware.

At first, the camera of worker interface captures the work area including work object and worker's arms. The captured images are sent to the instructor interface. At that time, the system corrects the work area image to overhead view for the tabletop display of instructor interface. Next, the instructor interface receives and the image displays it in the tabletop display. Instructors can make instructs such as gestures and pointing by fingers to the objects appeared on the display. RGB camera and depth sensor set above the tabletop display captures the instructions and sends them to the worker interface. Finally, projector of worker interface projects the image of instruction to the work area with the offset, which enables worker to work with interaction at hand. Now, RGB camera and depth sensor captures not only the instructions but also the image of work area indicated in the tabletop display because the camera and sensor are set downwards. Because of this, this system is like coupled mirror. To avoid it, instructor interface sends part of captured image upper to the tabletop display by using the depth information.

The previous work system adopted the method to display the image of instructor's arm. Hence it may be difficult to instruct finely when objects are small. The previous work proposed and implemented the scaling method of embodiment to solve that. When an instructor wants to instruct or see the work area finely, magnified image of the work area can be displayed in the display of the instructor interface.(Figure 2) This method makes an instructor see the work area in detail and instruct finely even to the small objects. When this method is used in the instructor interface, the instruction image is scaled down by the inverse scale of instructor's and projected to the work area. In addition, display range also moves to fit the interaction with real objects.

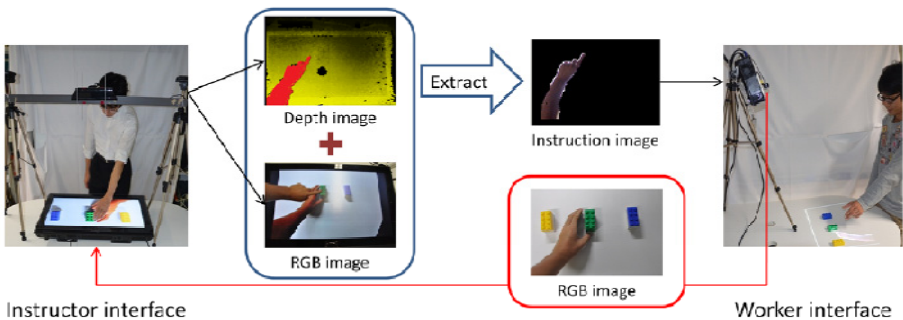


Fig. 1. Procedure of this system

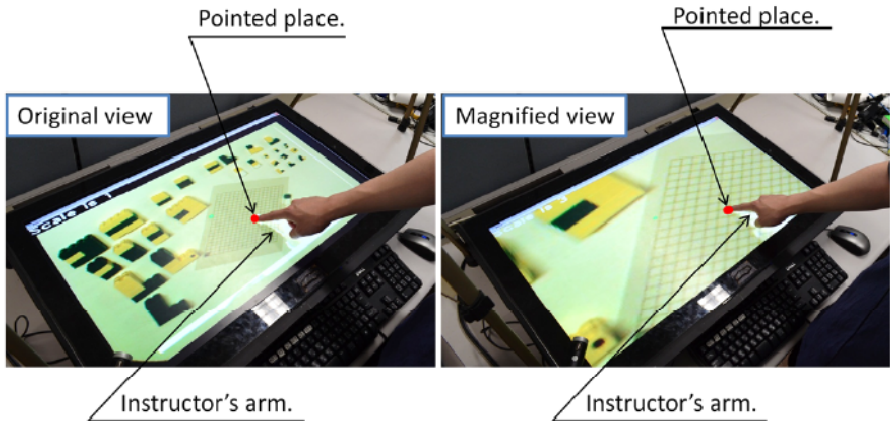


Fig. 2. Tabletop display in instructor interface (left: original view, right: magnified view)

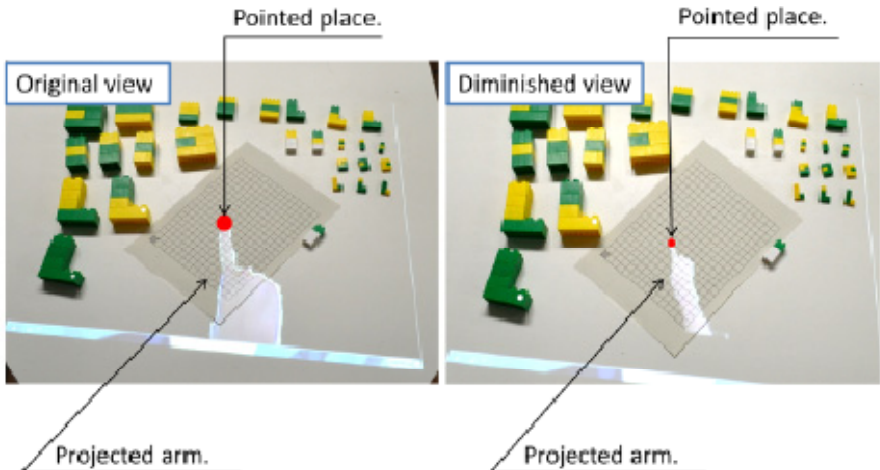


Fig. 3. Projected instructor's arm in worker interface (left: original view, right: diminished view)

3.2 User Study

This section describes a user experiment which was conducted to evaluate the effectiveness of the proposal method described in 3.1.

Instructor and worker interface were set on each remote place. Worker interface was set on the desk in the worker environment as shown in Figure 4. There were a grid paper including 16 x 20 cell which is 11.0 mm and 27 block cluster made of several three different sizes of blocks on the desk. Block cluster were placed around

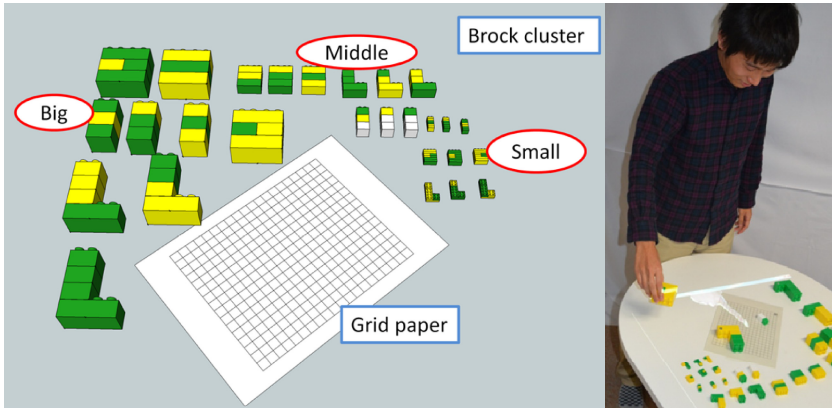


Fig. 4. The brock clusters in the worker environment and appearance of the worker experiment

the grid paper. Worker put the block cluster at the place indicated by instructor in the grid paper. In this experiment, worker place a block cluster on a grid paper by instructor layout plan. Six block clusters were used in a task. Several three different sizes of blocks were used. Each size of block cluster was used at least one block in a task. Block cluster was put on the grid paper to overlap a grid point with either edge of block cluster. Instructor provided direction according to the layout plan in order. Instructor provided direction watching the work environment displayed on the tabletop surface. Instruction was conducted by transmitting gestures and their voice.

1. Selecting a block cluster from the layout plan, and instruct the selected block cluster.
2. Indicating the angle of the block cluster on the grid paper.
3. Indicating the position of the either edge of block cluster and grid point by “pointing”.
4. Watching the position of the block cluster, replacing the right point.

After six block clusters were put on the grid paper, instructor makes sure of the put point. When the put point was correct, one task was completed. Instruction conditions were “scalable view condition” which was a proposal method and only original view. In “scalable view condition” condition, instructor could use the function magnifying image displayed on the tabletop display. In “Original view condition”, the function magnifying image was disabled during the experiment.

Subjects were able to select the scaling center by mouse click. Also, subjects were able to select the magnification percentage from x2.0 to x3.5 by keyboard. As well, keyboard and mouse are placed near the instructor interface. First, subjects conducted training tasks three times in “Scalable view condition” as a practice. After that, they conducted the actual tasks in “Scalable view condition” and “Original view condition” each three times. The order of instruction conditions was different for each subject to prevent the order effect. This experiment was conducted with twelve subjects (gender: twelve male; age: 22 to 28) who are not experienced this task as instructors

and one subject who has a good skill for this task as a worker (gender: male; age: 24). In this experiment, we measured the task completion time. After their tasks were ended, subjects answered the seven-level rating questionnaire whose contents were described below. Also, we let subject evaluate each size of blocks in “Scalable view condition” and “Original view condition”.

- Q1. Which condition do you transmit the instruction easier?
- Q2. Which condition do you think that worker can realize your direction easier?
- Q3. Which condition do you communicate to worker smoother?

3.3 Result

Figure 5 shows result of task completion time and questionnaire. “Scalable view condition” marks higher performances than original view condition”. Using the Wilcoxon signed rank test, there was significantly difference between “Scalable view condition” and original view condition ($p < 0.01$).

Using T-test (one-sample, test value=4), there were significant differences in all questions between “Scalable view condition” and original view condition ($p < 0.01$).

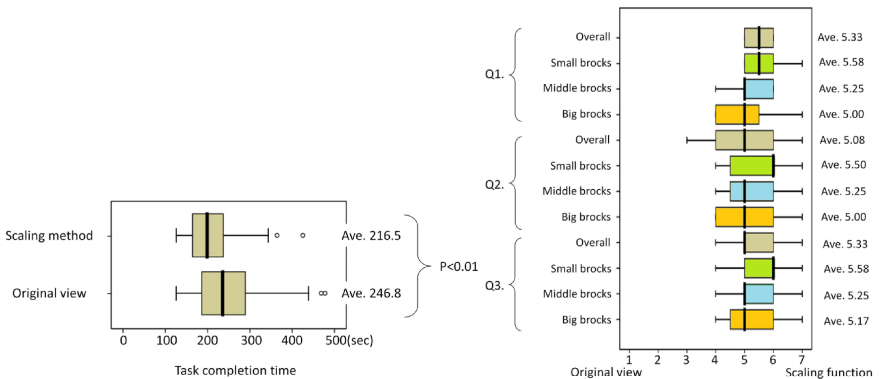


Fig. 5. Task completion time and result of questionnaire

4 Discussion of Previous Work

In previous work, result of user study just consists of task completion time and 3 items of questionnaires. Therefore, previous discussion [4] is not enough to argue the efficiency of the proposed method. For example, other criteria such as number of times of using scaling method are needed to argue that proposed method contributes to reducing task completion time because subjects were not forced to use the method. In addition, we also find some issues in questionnaire. In “Q1. the ease to instruct”, it is not appropriate to think the instruction of word and of gestures together. In “Q2. the

intelligibility of worker”, it is not appropriate for the instructor to judge it because it owes to only worker. To clear these issues, we observed closely the movie of user study and researched the behavior of instructor and worker with or without the scaling method.

In the result, all subjects instructed the right point with the scaling method of the embodiment. In addition, most subjects instructed the right point using the scaling method more times than not using the method. When subjects did not use the method, the right point is near the edge of the work area hence it is easy to instruct the right point even without the scaling method. Therefore, we confirmed that the proposed method contributed to the reducing task completion time. We discuss ”Q1. the ease to instruct”. We found the differences of the way of instruction between with and without scaling method only in the instruction of the right point. It is difficult to point just one vertex precisely without the scaling method because the size of instructor's finger is bigger than the size of square. Therefore, without the scaling method worker asked instructor to repeat the precise point to put the block cluster more times than with the scaling method, that affected the answer.

It is impossible to measure the ”Q2. the intelligibility of worker” precisely because it has been passing long after the experiment. Therefore, we judge it from the smoothness of communication between the instructor and the worker. It is reasonable to think that the intelligibility of worker is proportional to the smoothness of communication between the instructor and the worker and that the smoothness of communication between the instructor and the worker is proportional to the shortness of task completion time. Therefore, it is reasonable to think that the intelligibility of worker is proportional to the shortness of task completion time. With that, all the doubts of previous work are cleared.

However, by obtaining on the analyses of the movies, some problems appeared which previous work could not solve.

1. A block cluster put already hides the right point.
2. The block cluster putting now hides the right point.
3. Block clusters put already are moved incidentally, and instructor cannot make a smooth instruction.
4. It is hard to enlarge the image as the instructor supposes to put.
5. It is a fatigue to use the proposed method.

We discuss these problems. Problems (1) and (2) attributes to the occlusion of block clusters. In the user study, instructor cannot see areas just before block clusters and worker cannot see areas deployed block clusters because the camera of the worker interface faces the opposite direction of the instructor. View of the worker and the instructor are same if setting the camera of worker interface on the worker's side, but they cannot see areas deployed block clusters, either. If we try to avoid any occlusion, we should set more cameras or set a camera to the above the work area. However, this idea also has problems such that spaces to set cameras do not always exist and it force

worker to take more equipment. It is difficult to regard occlusion problems as the typical problems of this task because the more cubic task gets the more occlusion happen. Therefore, solution of this problem can be the guide for remote collaborative dealing with the cubic task.

Another problem of failure of communication can be seen at the same time of problem (2). We tell it in particular. Instructor sees the display to confirm if the block cluster is put on the right point because the cluster putting now hides the point. Instructor says nothing during confirming, that makes worker think that he put it at the wrong point and move it to the point that seems right. After moving it, instructor says that it was wrong and tells worker to move back to the right point. Failures of communication like this owe to the shortage of communications in some part, and owes to the impossibility of conjugate gaze in some part.

Problem (3) also happens when worker repeats the same job. We do not discuss this problem deeply because this scene is seen only without the scaling system and because remote collaborative work is not needed when repeating the same job because machine can take that place.

The one cause of problem (4) is that instructor cannot easily foresee the result of changing scale because scaling center and scale factor are needed to decide the display range. Higher scale factor makes it possible to instruct finely and makes it narrow the range of view. To satisfy fine instruction and wide range of view, scale factor should be taken continuous value different from this system that scale factor is chosen in the discrete 6 values.

Problem (5) is similar to problem (4), but we regard them as different issues. Problem (5) owes to the hardness of using the implemented scaling method. Instructor provides instruction by his own arms and hands, but he must use the keyboard and mouse which he does not use in normal instruction when using the scaling system. Moreover, he must move the scaling center to appropriate point in changing the point of focus. Therefore, it is necessary for instructor to move the scaling center by using scaling system that makes implemented method difficult to use. We propose the solution for problem (1) and (2) in this research.

4.1 Solution

When dealing with three-dimensional objects, problems (1)(2) which described in the previous section are commonly encountered. So, in remote collaborative work which deals with real object, the proposed method is implemented without additional equipment. Concretely speaking, save the image of process or initial state, by overlaying the current image and it, the area gotten behind can be checked.(Figure 5) It is considered that the proposed method is effective in situations such as the work area is hidden by the new installed objects. And, in order to examine the validity of the proposed method, perform the following experiments.

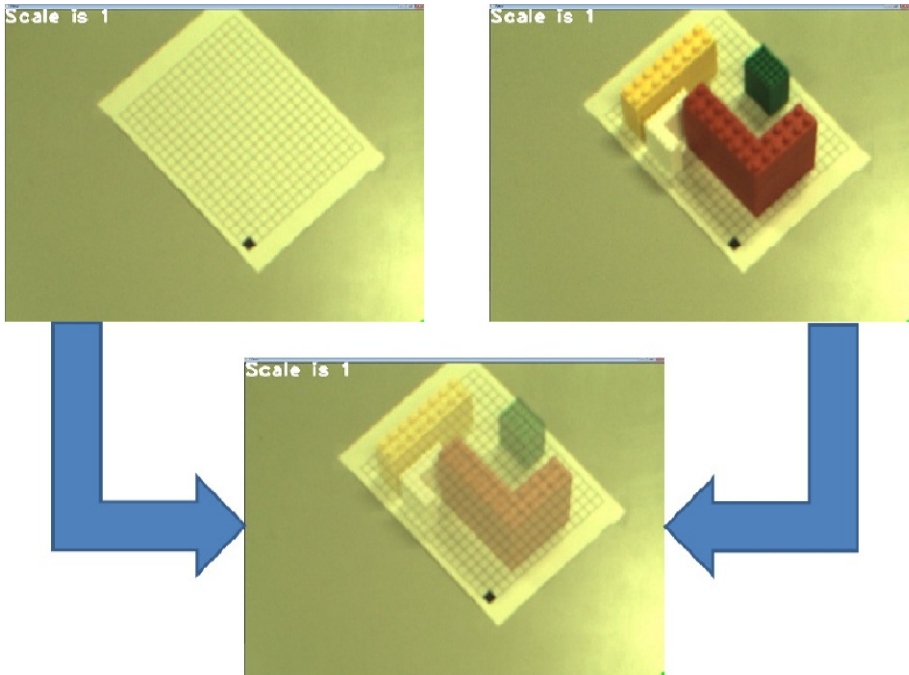


Fig. 6. Example of alpha-blending

5 Conclusion

We analyzed the movie of experiment of previous work again, then complemented the previous work and found some problems. Next, we proposed the method saving the past images and overlaying the past image to the current image and evaluated it. We suppose to conduct user study to verify the efficiency of proposed method in a cubic task and improve the system by reference to the result of user study.

References

1. Spatial workspace collaboration: A sharedview video support system for remote collaboration capability. In: Proc. CHI 1992, pp. 533–540 (1992)
2. Fussell, Setlock, L.D., Kraut, R.E.: Effects of head-mounted and scene-oriented video systems on remote collaboration on physical tasks. In: Proc. CHI 2003, pp. 513–520 (2003)
3. Kraut, R.E., Miller, M.D., Siegal, J.: Collaboration in performance of physical tasks: Effects on outcomes and communication. In: Proc. CSCW 1996, pp. 57–66 (1996)
4. Uemura, K., Sakata, N., Nishida, S.: Improving Visibility of Gesture Image with Scaling Function for Tabletop Interface in Remote Collaboration. Transactions of the Virtual Reality Society of Japan 17(3) (2012) (in Japanese)

5. Mikawa, M., Matsumoto, M.: Smooth and easy telecommunication using CTerm. In: Proceedings of IEEE SMC 1999, pp. 732–737 (1999)
6. Yamazaki, K., Yamazaki, A., Kuzuoka, H., Oyama, S., Kato, H., Suzuki, H., Miki, H.: In: Proceedings of the Sixth Conference on European Conference on Computer Supported Cooperative Work, pp. 239–258 (1999)
7. Sakata, N., Kurata, T., Kato, T., Kouroggi, M., Kuzuoka, H.: WACL: Supporting telecommunications using wearable active camera with laser pointer. In: 2003 Proceedings. Seventh IEEE International Symposium on Wearable Computers, pp. 53–56 (2003)
8. Kuzuoka, H., Furusawa, Y., Kobayashi, N., Yamazaki, K.: Effect on Displaying a Remote Operator's Face on a Media Robot. In: Proceedings of ICCAS 2007, pp. 758–761 (2007)
9. Yamamoto, Xu, H., Sato, K.: Palmbit-silhouette: Desktop accessing by superimposed silhouette of the palm. In: Interaction 2008, pp. 109–116 (March 2008) (in Japanese)
10. Kondo, Kurosaki, K., Iizuka, H., Ando, H., Maeda, T.: View sharing system for motion transmission. In: Proceedings of the 2nd Augmented Human International Conference (March 2011)
11. Kirk, D., Crabtree, A., Rodden, T.: Ways of the hands. In: Proc. 9th European Conference on Computer-Supported Cooperative Work, France, pp. 1–21 (September 2005)
12. Tang, A., Pahud, M., Inkpen, K., Benko, H., Tang, J.C., Buxton, B.: Three's company: understanding communication channels in three-way distributed collaboration. In: Proc. ACM Conference on Computer Supported Cooperative Work, Savannah, USA, pp. 271–280 (February 2010)
13. Yamashita, Kuzuoka, H., Hirata, K., Aoyagi, S., Shirai, Y.: Supporting fluid tabletop collaboration across distances. In: Proc. Annual Conference on Human Factors in Computing Systems, Vancouver, Canada, pp. 2827–2836 (May 2011)
14. Izadi, A., Agarwal, A., Criminisi, J., Winn, A., Blake, A., Fitzgibbon, A.: C-Slate: A Multi-Touch and Object Recognition System for Remote Collaboration using Horizontal Surfaces. In: IEEE Workshop on Horizontal Interactive Human Computer Systems, Rhode Island, USA, pp. 3–10 (October 2007)

Two Handed Mid-Air Gestural HCI: Point + Command

Matthias Schwaller, Simon Brunner, and Denis Lalanne

Department of Informatics, University of Fribourg, 1700 Fribourg, Switzerland
firstname.lastname@unifr.ch,
brunner.simon@gmail.com

Abstract. This paper presents work aimed at developing and evaluating various two-handed mid-air gestures to operate a computer accurately and with little effort. The main idea driving the design of these gestures is that one hand is used for pointing, and the other hand for four standard commands: selection, drag & drop, rotation and zoom. Two chosen gesture vocabularies are compared in a user evaluation. The paper further presents a novel evaluation methodology and the application developed to evaluate the four commands first separately and then together. In our user evaluation, we found significant differences for the rotation and zooming gestures. The iconic gesture vocabulary had better performance and was better rated by the users than the technological gesture vocabulary.

Keywords: Gestural interfaces, Two-hand interaction, User evaluation.

1 Introduction

Gestural user interfaces have become omnipresent in recent years. At the time of writing, the Microsoft Kinect device is one of the most used devices to recognize hands-free gestures. The great success of the Kinect device is mostly due to its low price and its accuracy for body tracking. The device comes with an RGB camera and a depth sensor, both with a resolution of about $640 * 480$ pixels and with a frame rate of 30 fps.

Until now, most hands-free gestures required big arm movements to recognize them. This paper aims at studying subtle two handed gestures in order to perform basic gestural interaction to operate standard programs on a personal computer. We decided to use a two handed paradigm with one hand for pointing and the second hand for the gestural interaction because (1) most of the users in Nancel et al. [1] rated two handed gestures as less tiring than one handed gestures, (2) there are more possibilities for different commands and (3) the gesture cannot influence the pointing accuracy (for instance a selection) and it further allows moving the pointer while a command is done.

Two different sets of gestures have been designed and compared through a user evaluation. Since there is no standard methodology to compare gestures, other than for a selection of gestures, we compared the 3 other gestures with small repetitive tasks consisting of manipulating geometrical objects. Both gesture vocabularies are

composed of four commands: selection, drag & drop, rotation and zoom. The drag & drop command is composed of a selection and a release. The first gesture vocabulary is technology driven, meaning that this vocabulary is designed to guarantee robust gesture recognition, while the second gesture vocabulary is user centered and is designed to have easy-to-use and easy to remember gestures.

The presented evaluation application permits to evaluate the four commands (selection, drag & drop, rotate and zoom) first separately and then together. The selection command evaluation is based on part 9 of the ISO 9241 standard for non-keyboard input devices and Fitt's law (multi-directional tapping test). For the other commands geometric objects have to be moved, rotated and/or resized. This gesture evaluation application is another contribution of this paper.

The remainder of the paper is structured as follows: First, we give an overview of some related work and then present the two gesture vocabularies. Furthermore, we present the test application and the evaluation with its various results. Finally we discuss the results, articulate a conclusion and discuss future work.

2 Related Work

Various research work has been done in the past concerning the recognition of two handed gesture input. Nancel et al. [1] presented an evaluation where they compared uni- vs. bimanual interaction. In their study they found that two handed techniques were less tiring than one handed techniques. They used the same paradigm as we do, namely to use the right hand for pointing and the left hand for gestural interactions. In their study they found that 3D in the air gestures are more tiring than gestures with either a 2D surface or a 1D path. Therefore we created more subtle gestures which require less movement and should therefore be less tiring. In their paper they also presented an interesting approach to evaluate moving and zooming of virtual circles. We use a similar approach to evaluate the zoom and the rotate gestures.

A Kinect as a sensor for two handed interaction was also used by Gallo et al. [2]. They used the same selection gesture as we use in the iconic vocabulary. For the rotation and the zooming they used both hands. We, on the contrary, use only one hand for commands and the other hand for pointing to the object we would like to modify.

A bimanual handle bar metaphor was introduced by Song et al. [3]. In their implementation they used two pointing gestures to position a handle bar, which is in fact a virtual stick that traverses the virtual object. Further, to manipulate the objects two fists were used to stretch or rotate this handle bar which zoomed or rotated the objects. Since we use only one hand for the commands and our gestures need less movement, our gestures therefore need less effort.

Ziegelbaum et al. [4] did some important work on two handed interaction. Their work was built on the g-speak by Oblong¹. For their rotation and translation they used two pinch gestures (thumb and index finger touching). The movement which the user needs to do to generate commands is bigger than in our work. Furthermore, for their system the user has to stand, while for our system the user can sit in a chair with armrests.

¹ <http://www.oblong.com/>

The WUW which is also known as Sixthsense was introduced by Mistry et al. [5] and proposed two handed zoom in and zoom out gestures very similar to the zoom gesture presented by Ziegelbaum et al. [4]. Their system is wearable and therefore the user has to wear a webcam on his chest.

The two handed zoom and rotation gestures presented by Kamel Boulos et al. [6]. are used for navigating in Google Earth. To navigate the two hands have to make fists the use is then very similar to using Google Earth with a touch screen. Instead of touching the screen with one finger, the user has to make a fist and can thus zoom and rotate with two fists. Other pan and zoom gestures to navigate in Google Earth were presented by Stellmach et al. [7]. In their study they compared 4 pan and zoom alternatives. They found that the gestures, in which the non-dominant hand had to be held up while performing the gestures, were faster than two handed gestures where both hands interacted for zooming, although the user commented that it was painful to constantly hold up their hand. For our gestures, the gestural command is done with the left hand, while the right hand is pointing to the object which has to be manipulated. This gesture is therefore somehow close to their gesture with the difference that our right hand does not have to be help up over the shoulder but is pointing in a more comfortable position.

Another two handed scenario was introduced by Gustafson et al. [8] and called “Imaginary Interfaces”. In their approach they used the left hand to show in the air where the user wants to draw and the right hand was used for drawing. They have a similar gesture paradigm because in their system, as in ours, the right hand is used for pointing / drawing. The difference is that with their right hand they indicate not only the position, but also if the hand is pointing or not (with a pinch gesture) while in our system the right hand indicates only the position, which is cognitively simpler for users: one hand used for pointing, the other for commands.

3 Design of Gestures

During this research we developed about 20 gestures classified into two groups of gestures, although most of the gestures are improvements of others. The first group of gestures is called technological and is designed to ensure robust and fast gesture recognition. The second gesture group, called iconic, was built to be ergonomic and easier to memorize by users. Two different sets of gestures were finally chosen from the 20 gestures and compared through a user evaluation. For both vocabularies the left hand is used for the commands and the right hand for the pointing (see Fig. 1).

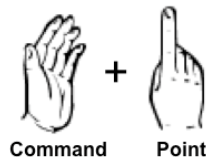


Fig. 1. Point + Command two handed gestural interaction paradigm

For the finger detection the Candescent NUI library² was used. This library gives the number of recognized fingers per hand as well as their location (fingertip), the volume of the hand, the palm position etc. For the pointing, which is done with the right hand, the highest point is used. Due to this fact, the hand can have any desired posture, which helps to reduce fatigue, although for the user it is easier to have only one finger outstretched. This prevents the cursor from switching from one finger to another if the user makes small rotations of the hand while he or she moves the cursor. In both gesture vocabularies the rotations and the zooms are continuous and thus clutch free. This means that the hand does not have to be repositioned during the gesture.

The rotation and the zoom are implemented in such a way that the center of the rotation and the focus point for zooming are always the middle point of the figure (marked with a blue dot, see **Fig. 6**).

3.1 Technological Gesture Vocabulary

The first gesture vocabulary is called technological and is designed to facilitate the efficiency and reliability of the gesture recognition. For this gesture set, sliding movements are used for the rotation and zoom. In order to activate the two actions, only one finger has to be present (see Fig. 2). As soon as only one finger is detected by the system, a red square appears on the screen with arrows and the names of the actions (Fig. 2). It serves as a neutral zone to prevent undesired action in the system. To execute an action (rotation or zoom) the finger has to go outside the red square. The two actions, rotation and zoom, cannot be done together.

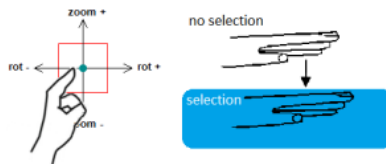


Fig. 2. The technological gesture set; on the left side rotation and zoom and on the right side selection

The user has to move the finger horizontally to rotate an object (see Fig. 3-b); left to rotate left, and right to rotate right. To perform a zoom the user has to move the finger vertically (see Fig. 3-c); up to zoom in and down to zoom out. The selection is done by showing the entire hand and making a vertical movement downwards, if the hand is below a certain threshold, shown with a blue rectangle (see Fig. 3-d), a selection is performed and if the hand moves up, a release is performed.

This gesture vocabulary requires only two postures (showing one finger and showing all fingers) for the left hand which supports robust gesture recognition and due to the small number of postures, it is also easy to remember for the user.

² <http://candescentnui.codeplex.com/>

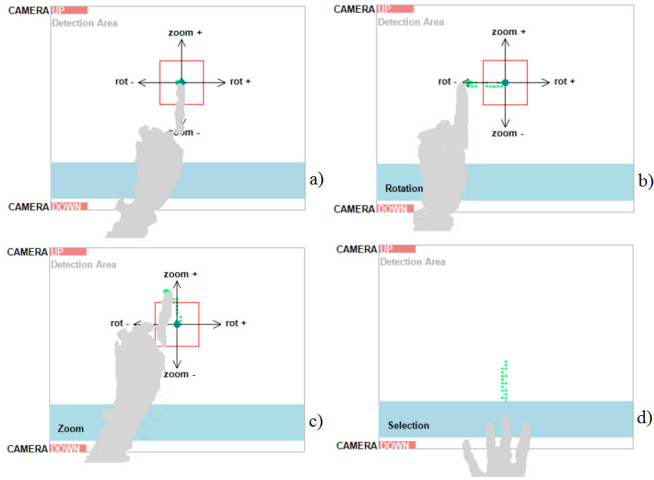


Fig. 3. Feedback technological gesture vocabulary

3.2 Iconic Gesture Vocabulary

For the second gesture set we followed a different approach than we did for the first gesture set and designed the gestures to be more natural and easier to memorize by users. In this iconic gesture vocabulary the rotation and the zoom are operated with a circular movement (see Fig. 4). For doing a rotation, only one finger has to be outstretched. For a zoom, two fingers must be outstretched. As feedback, a line is drawn between the activation point (darker green in Figure 5) and the current point (brighter green in Figure 5).

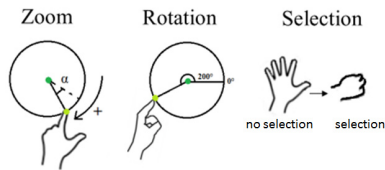


Fig. 4. The iconic gesture set; on the left side the zoom gesture, in the middle the rotation gesture and on the right side the selection gesture

The rotation gesture contains four parts. The first part is for the activation of the gesture, which is done by hiding all but one finger. At this point a dark green point is shown in the feedback (see dark green point in Figure 5 a). In the second part the user has to move the hand in one direction. Now, a violet line is drawn in the feedback (see Figure 5 a). As a third step the user has to move his hand around the green middle point. To stop the gesture, which is the fourth step, the user has to stretch out all fingers. The angle between the activation and the end is the angle which is used to rotate the object. In fact the object is already turning while the gesture is performed. If the user moves with a small radius around the center point (initial point), the rotation

is fast but the user has less precision. Otherwise, if the user moves his hand further outside the center, the object turns more slowly and the user thus has more precision.

The difference between the zoom and the rotation gestures is the posture, which the user has to make in order to activate the gesture. For the zoom gesture, the user has to have two fingers outstretched (see Figure 5 b). The angle of this technique obviously cannot be mapped directly to the zoom level. Therefore we compare the actual angle to the angle before and then do either a small zoom in or zoom out.

The selection in this gesture vocabulary works like grabbing an object with the hand. For a selection the user has to make a fist (see Figure 5-c) and for a release the user has to reopen the hand again.

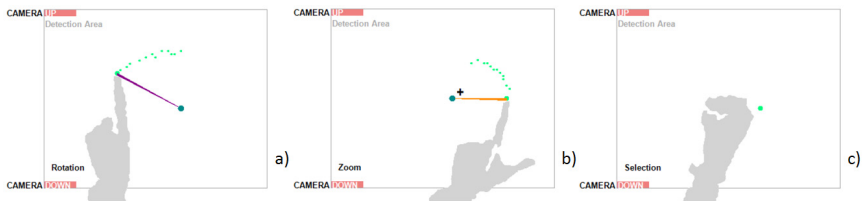


Fig. 5. Feedback iconic gesture vocabulary

4 Test Application

The developed evaluation application permits to evaluate the four commands (se-lection, drag & drop, rotate and zoom) first separately and then together (see Fig. 6). Before the five exercises (the four commands separately and then together) each user has to do a training session to make sure that they have fully understood how to perform the gestures (see Fig. 6-a). The training takes a minimum of 90 seconds but can be longer if the user needs more training. Before each task there is a small animation displayed on the screen in which the task is shown to the user.

In the first level, which is the selection level (see Fig. 6-b), the user has to click on round targets. There are 12 round targets arranged on a circle. As soon as the first circle is clicked the 2nd will be highlighted and so on, as proposed in part 9 of the ISO 9241 standard for non-keyboard input devices (multi-directional tapping test). In the second level, which is the drag & drop level, the user has to move a circle into a zone. In Fig. 6-c the orange circle has to be moved into the blue rectangle. The task is validated when the blue middle point of the circle is within the rectangle. In the rotation level, which is shown in Fig. 6-d, several geometrical objects have to be rotated in order to fit the target which is represented by a border. Since the object only has to be rotated, the “figure to rotate” and the final target have the same center point. In order to make the task possible to accomplish, a small tolerance is allowed and the target turns green when the task is accomplished. The resizing level is shown in Fig. 6-e. This level acts very similarly to the rotation level. The final target is shown by the border of the figure and the center points are the same. Also for this level, there is a small tolerance which is allowed and the target turns green when the task is

accomplished. The final level is shown in Fig. 6-f. This final level combines the selection, drag & drop, rotation and resizing. The user has to select a geometrical object, move it to the right place, turn it to the correct angle and zoom it to the correct size. These actions can be done in any desired order.

To configure the application an xml file is used. The file contains all the pieces with their location, size and angle. Further it is possible to configure whether the piece can be zoomed, rotated, selected and moved. In the configuration file, the final location, size and orientation of pieces (used for the levels rotation, resize and final) are also specified.

While doing the different commands with the left hand, the right hand is used for pointing. As pointer feedback we use a cross-hair feedback i.e. small red circle with a cross (see Fig. 6 a-f). The selection command evaluation is based on part 9 of the ISO 9241 standard for non-keyboard input devices and Fitt's law (multi-directional tapping test).

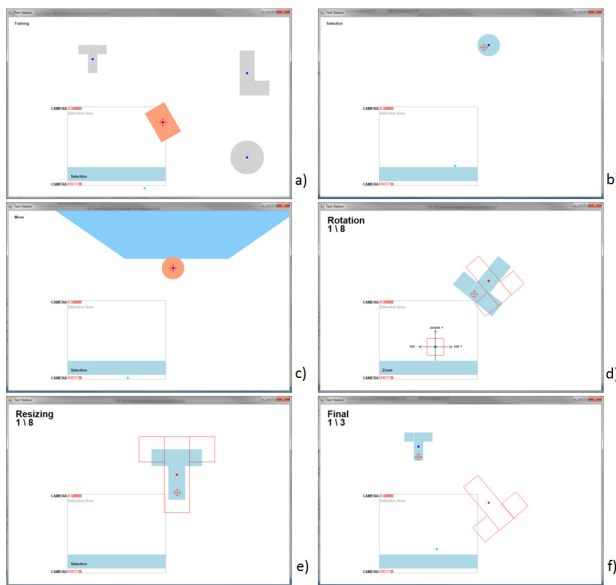


Fig. 6. The test application: the training panel (a) and the 5 exercises (b-f)

The objects in the test application can either be gray (normal state), blue (cursor is over it) or orange (selected). For the target objects, only the border is shown in red. If an object reaches the target, it becomes green and a message is displayed (see Fig. 7).

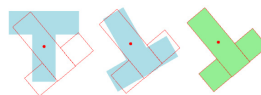


Fig. 7. Feedback geometrical objects

5 Evaluation and Results

We conducted a within subject user evaluation with 10 users to compare the two gesture vocabularies. The comparison of the gesture vocabularies was done by comparing the time needed by the users to accomplish the tasks. In our evaluation all the users tested both gesture vocabularies. The gesture vocabularies were counter balanced to reduce carry over effects (fatigue and learning). For each gesture vocabulary the users had to select 12 objects as well as to move 12 objects. Further, the user had to rotate 8 objects and resize 8 objects. In the final level, the users had to move, rotate and resize 3 objects. The evaluation was done on a personal computer with an office chair with arm rests. Several paired samples t-tests were conducted. Significant difference was found between the two rotation gestures: $t(9) = 3.72$, $p = .0048$. Significant difference was also found between the two resize gestures: $t(9) = 2.57$, $p = .03$. For the other pairs there was no significant difference. The results are presented in Table 1.

Table 1. Results of the user evaluation

		mean	median	low	high	std dev	t-value	df	t-test	
Selection	techno.	4.621	4.425	2.95	6.99	1.32	0.737	9	0.48	----
	iconic	4.26	4.184	3.32	6.52	0.915				
Move	techno.	5.475	4.961	3.83	8.91	1.44	1.118	9	0.266	----
	iconic	4.87	4.852	3.92	6.97	0.89				
Rotation	techno.	17.57	16.18	9.14	28.7	5.69	3.72	9	0.0048	**
	iconic	14.07	14.07	9.18	22.1	4.13				
Resizing	techno.	14.63	14.61	9.61	19.8	2.79	2.57	9	0.03	*
	iconic	12.28	12.06	9.16	15.7	1.86				
Final	techno.	54.3	54.74	32.5	69.1	12.3	-0.25	9	0.808	----
	iconic	55.34	56.96	31.8	73.3	12.1				

null hypothesis significance level: * = < 0.05; ** = < 0.01; *** = < 0.001; * = < 0.05

After the first gesture set, as well as after the second gesture set, the users had to fill out a questionnaire, based on the one proposed in ISO 9241 part 9 annex C. The questions have a Likert scale from 1 to 7, where 7 is the best. For the fatigue questions it is the inverse, i.e. 1 is the best. This questionnaire permitted to measure user satisfaction as well as fatigue. The results of the questionnaire are presented in Table 2.

Table 2. Results of the questionnaire based on the ISO standard

	Smoothness during operations	Effort required during operations	Accuracy	Rotation ease of use	Zoom ease of use	Selection ease of use	Operation Speed	General comfort	Feedback quality	Overall quality	Finger fatigue	Wrist fatigue	Arm fatigue	Shoulder fatigue	Back fatigue	Overall fatigue
Techno	5.4	2.5	4.4	4.2	4.7	4.4	5.1	5.0	5.9	5.2	1.8	1.3	5.4	2.5	1.4	3.8
Iconic	5.1	3.0	5.2	4.3	4.9	6.5	5.2	4.4	5.1	5.4	2.3	1.3	5.3	2.6	1.5	4.0

To analyze the perceived quality of the two gesture vocabularies we conducted Wilcoxon signed rank tests, since the data is not normally distributed. We found only two significant differences for “Selection ease of use” and “feedback quality”. In general the commands are slightly better rated for the iconic gesture vocabulary but only the selection is significantly better. For the fatigue the technological gesture set is rated slightly better but not significantly. The general comfort seems to be better for the technological vocabulary, although not significantly. This could be influenced by the fatigue. The second significant difference is the feedback which was better rated for the technological gesture vocabulary.

There is a reason why the fatigue was rated worse for the iconic gestures, namely that the users have to switch the commands for zooming and rotating in the iconic gestures set (two different activations; one and two fingers) whereas for the technological gesture set the activation gesture (one finger) is the same for both and thus it is faster to change between those gestures.

6 Conclusion

In this article we presented two gestural vocabularies with a 2 handed point & command paradigm. The first one is based on a technological approach which is easier to detect. The second one is based on an iconic approach and is therefore more natural to use. Secondly, the paper presented a test application on which the gestures were compared. Finally a user evaluation was presented.

By analyzing the results of user performance in the user evaluation we found significant differences for the rotation and the resize gestures. Overall, the iconic gesture vocabulary had slightly better performance. Concerning the user perception, users preferred the selection of the iconic gesture vocabulary, which is more natural (grabbing an object), over the more technological one. For the other commands there was no significant difference. In this user evaluation we did not have the opportunity to check how easily the gestures were remembered. Analyzing the questionnaire and the comments of the users during the evaluation, we suppose that the iconic gesture set is easier to remember. For those reasons in future work we suggest to develop more gestures which are more natural to perform than gestures which are easy to recognize, in order to augment the acceptance of users as well as the easiness of performing the gestures.

We plan to develop other gestures which would be more subtle but still easy to perform and easy to remember by users, but which would require machine learning because they would be too subtle for stable recognition otherwise. In a second step we would then compare the machine learning gestures with the ones proposed in this paper. For this comparison we would like to make two evaluations, the first one with the application proposed in this paper and the second one with another application such as Google Earth. In the future we would also like to compare two handed gestures with the paradigm where one hand is used for pointing and one hand for the gesture command, versus one handed gestures, since then both hands could do

gestures at the same time. We also plan to design gestures in combination with speech input and would like to compare those with the two handed gesture paradigm presented in this paper.

Acknowledgment. Grateful acknowledgement for proofreading and correcting the English goes to Agnes Lisowska Masson.

References

1. Nancel, M., Wagner, J., Pietriga, E., Chapuis, O., Mackay, W.: Mid-air pan-and-zoom on wall-sized displays. In: Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems, CHI 2011, p. 177. ACM Press, New York (2011)
2. Gallo, L., Placitelli, A., Ciampi, M.: Controller-free exploration of medical image data: experiencing the Kinect. *Computer-Based Medical* (2011)
3. Song, P., Goh, W.B., Hutama, W., Fu, C.-W., Liu, X.: A handle bar metaphor for virtual object manipulation with mid-air interaction. In: Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems - CHI 2012, p. 1297 (2012)
4. Zigelbaum, J., Browning, A., Leithinger, D., Bau, O., Ishii, H.: g-stalt: a chirocentric, spatiotemporal, and telekinetic gestural interface. In: Proceedings of the Fourth International Conference on Tangible, Embedded, and Embodied Interaction - TEI 2010, p. 261. ACM Press, New York (2010)
5. Mistry, P., Maes, P., Chang, L.: WUW - Wear Ur World - A Wearable Gestural Interface. In: Proceedings of the 27th International Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA 2009, p. 4111. ACM Press, New York (2009)
6. Boulos, M.N.K., Blanchard, B.J., Walker, C., Montero, J., Tripathy, A., Gutierrez-Osuna, R.: Web GIS in practice X: a Microsoft Kinect natural user interface for Google Earth navigation. *International Journal of Health Geographics* 10, 45 (2011)
7. Stellmach, S., Jüttner, M., Nywelt, C., Schneider, J., Dachselt, R.: Investigating Freehand Pan and Zoom. In: Reiterer, H., Deussen, O. (eds.) *Mensch & Computer 2012: Interaktiv Informiert – Allgegenwärtig und Allumfassend!?*, pp. 303–312. Oldenbourg Verlag, München (2012)
8. Gustafson, S., Bierwirth, D., Baudisch, P.: Imaginary Interfaces: Spatial Interaction with Empty Hands and without Visual Feedback. In: Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology - UIST 2010, p. 3. ACM Press, New York (2010)

Experimental Study Toward Modeling of the Uncanny Valley Based on Eye Movements on Human/Non-human Faces

Yoshimasa Tawatsuji¹, Kazuaki Kojima², and Tatsunori Matsui²

¹ Graduate School of Human Sciences, Waseda University

² Faculty of Human Sciences, Waseda University

2-579-15, Mikajima Tokorozawa Saitama, 359-1192, Japan

wats-kkoreverfay@akane.waseda.jp, koj@aoni.waseda.jp,
matsui-t@waseda.jp

Abstract. In the research field of human-agent interaction, it is a crucial issue to clarify the effect of agent appearances on human impressions. The uncanny valley is one crucial topic. We hypothesize that people can perceive a human-like agent as human at an earlier stage in interaction even if they finally notice it as non-human and such contradictory perceptions are related to the uncanny valley. We conducted an experiment where participants were asked to judge whether faces presented on a PC monitor were human or not. The faces were a doll, a CG-modeled human image fairly similar to real human, an android robot, another image highly similar and a person. Eyes of the participants were recorded during watching the faces and changes in observing the faces were studied. The results indicate that eye data did not initially differ between the person and CG fairly similar, whereas differences emerged after several seconds. We then proposed a model of the uncanny valley based on dual pathway of emotion.

Keywords: The uncanny valley, eye movements, dual pathway of emotion, humanlike agent.

1 Introduction

In the research field of human agent interaction, it is a crucial issue to clarify the effect of appearance of a robot in the real world or a virtual character on a PC monitor on human impressions toward it [1]. Here, we refer to such a robot or character as an *agent*. *The uncanny valley*, which was coined by Mori [2], is one critical topic in considering appropriate appearances of agents in terms of impressions from people. It hypothesized that although human familiarity toward an agent increases as an appearance of the agent gets more humanlike, it drastically decreases to the bottom of the valley when the agent-appearance is almost the same to but slightly different from real human as illustrated in Fig.1. Several studies have approached to what causes the phenomenon hypothesized in the uncanny valley and how, even though its mental mechanism has not been yet clarified.

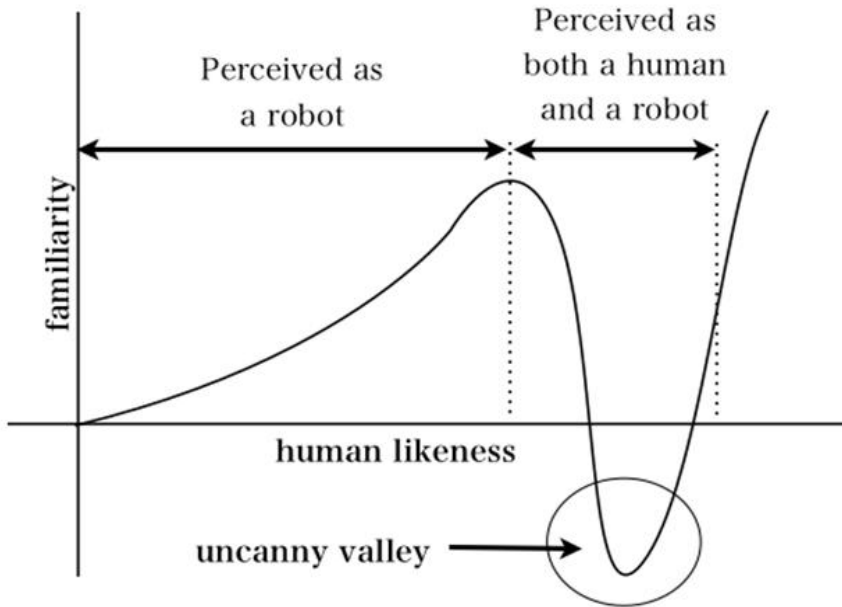


Fig. 1. Basic concept of the uncanny valley (partially altered from that in [2])

Some studies provided empirical data regarding how people observed a humanlike agent and how observation of an agent differed from that of a person. Noma et al. [3] revealed that about 75% of participants recognized a moving humanlike robot as human when observing it in short time, whereas they noticed that it was not human as time passed. These facts must imply that people can perceive a humanlike agent as human in a short interaction and that people can simultaneously recognize such an agent as both human and non-human. Minato et al. [4] proved that perceptual responses to a human and a humanlike agent were different through an experiment where participants had a conversation with a person or an android highly similar to human. Behaviors of the participants were observed by recording eye fixations. They reported that the proportion of eye fixations on the person's face was smaller than that on the android's face. This result indicates that perceptual processes of human and non-human differ from each other. From these studies, we hypothesized that people can perceive an agent as human at an earlier stage in interaction even if they finally notice it as non-human and such contradictory perception are related to the uncanny valley. According to findings by Minato et al., changes of perceptions when facing an agent can be verified by obtaining data of eyes. Our hypothesis predicted that eye movements of those who observed a person or an agent extremely similar to real human did not differ initially, whereas differences between them emerged after observing for a while.

This study experimentally investigated changes in human perceptual processes of a person and humanlike agents. We conducted an experiment where participants judged whether each of the person or agents was human or non-human. Their perceptual

processes were studied by recording and analyzing their eye movements. We then discussed perceptual processes of humanlike agents based on the experimental results and modeled the processes to provide explanation regarding how the uncanny valley occurs.

2 Experiment

2.1 Method

This experiment used five pictures of faces of (a) a doll, (b) a CG-modeled human image fairly similar to real human, (c) an android robot, (d) another image highly similar, and (e) a person as shown in Fig. 2. These faces were selected from several web pages to present faces whose similarities to real human got gradually higher.

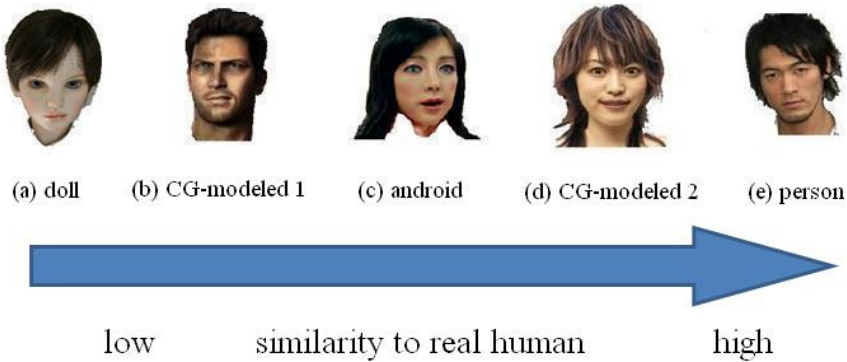


Fig. 2. Five faces used in experiment

In our experiment, participants were asked to judge whether each of the faces presented on a PC monitor was human or not. Each face was located at the center of the monitor. To control the initial location of eye of the participants, a white page where a cross was depicted at its central point was inserted before presenting each face. Eyes of the participants were recorded during watching the faces and eye fixations on the faces were estimated with EMR-AT VOXER produced by nac Image Technology. The participants were told that each face was presented for one minute and asked to write their judgments on a paper sheet. The faces were presented in the order of the doll, CG-modeled 2, android, person, and CG-modeled 1.

Some of the participants were asked to respond to two questionnaires regarding the faces after the judging task. The questionnaires included (Q1) *how difficult was the judgments of each face?* and (Q2) *which parts of faces did you pay attention to when judging?* The participants responded to Q1 on a three-point scale where 1 denoted easy and 3 denoted difficult.

2.2 Data Analysis

According to Yarbus [5], people frequently gaze at a region including the eyes, nose and mouth during watching at a person's face. These facial areas are important for human to seize some social information about others. Thus, we calculated a length of time when each area was gazed at for each face.

We used dFactory, analysis software for eye movement data, to calculate how long the participants gazed at each facial area. The calculation was conducted in three steps. First, the screen of the monitor was divided into 16 x 16 small blocks. Second, areas denoting the right eye, left eye, nose and mouth were defined. Each area comprised a block of the respective face part and its surrounding blocks. For example, the right-eye area comprised a block including the center of the pupil of the right eye and eight blocks surrounding the pupil block. Fig. 3 indicates the four areas in case of the CG-modeled 1. Finally, total time length of eye fixations on each area was calculated by adding time of eye fixations on each of comprised blocks. To confirm changes in perceptual processes, the analysis of eye-fixation time was performed in three time spans: 5 seconds, 10 seconds and 30 seconds from the start of face presentation.

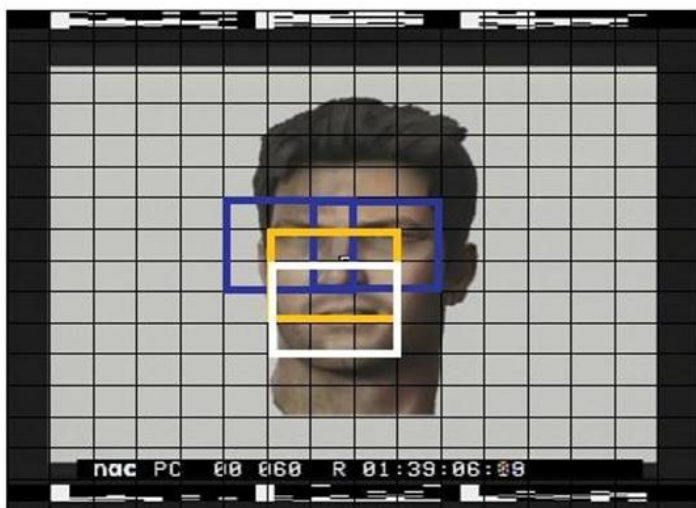


Fig. 3. Areas of right eye, left eye, nose and mouth

3 Results

Twenty one undergraduates (18 males and 3 females) participated in the experiment.

3.1 Judgment of Human/Non-human to Each Face

The proportions of participants who judged each face as human were as 28.6% for the doll, 19.1% for the CG-modeled 1, 19.1% for the android, 90.5% for the CG-modeled

2 and 100.0% for the person. Fig.4 indicates the proportions of judgments. Those results were mostly corresponding to our assumption of the similarities to real human. The android and CG-modeled 1 can be considered as the most unsimilar. Although the doll was more similar than the two, it was also evaluated as less humanlike. The CG-modeled 2 was the most humanlike, and the person was correctly judged as human. To precisely study differences in observing human and non-human faces, we analyzed eye data of the CG-modeled 1 which was mostly judged as non-human, that of the CG-modeled 2 which was judged as human although it was actually non-human and that of the person.

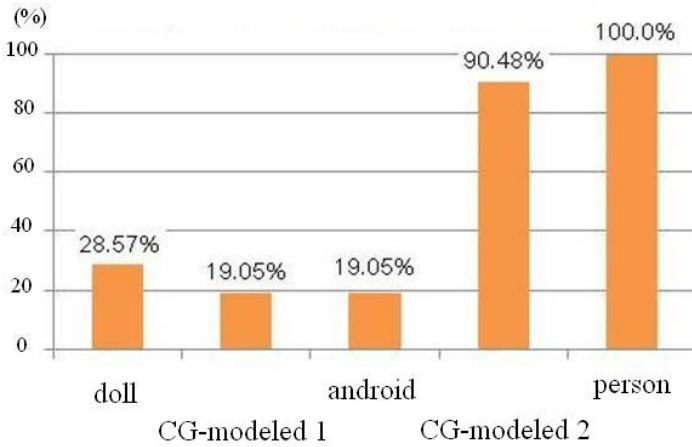


Fig. 4. Proportions of participants who judged each face was human



Fig. 5. Transactions of eye fixations of a participant in initial five seconds

3.2 Time of Eye Fixations on Areas of Each Face

Gaze data of 15 participants who judged the CG-modeled 2 as human and the CG-modeled 1 as non-human was analyzed. However, data of 7 participants was excluded due to its poor quality. Thus, data of the other 8 was actually used. Fig.5 shows examples of transactions of eye fixations during observing each face in the initial five seconds. The size of each circle denotes the length of total time of eye fixations at the respective point.

Table 1 indicates averages of time length of eye fixations on the four areas of each face in each time span. The t-test revealed significant differences of time length of eye fixations on the right eye areas among the three faces. Fig.6 shows average time of eye fixations on the right eye area of each face in each time span. In initial 5 seconds, eye fixations on the right eye area of the CG-modeled 2 was significantly longer than that of the CG-modeled 1 ($p<.01$) and that of the person ($p<.01$). In initial 10 seconds, eye fixations on the right eye area of the CG-modeled 2 was significantly longer than that of the CG-modeled 1 ($p<.01$) and that of the person ($p<.01$), and the difference between the CG-modeled 1 and person was moderately significant ($p<.10$). In entire 30 seconds, the participants observed the right eye area of the CG-modeled 2 more frequently than that of the person ($p<.01$), and the difference between the CG-modeled 1 and person was moderately significant ($p<.10$).

Table 1. Average of time length of eye fixations on left eye area, right eye area, nose area and mouth area of each face (sec.)

5 seconds	CG-modeled 2	CG-modeled 1	Human
On left eye	2.496	2.580	2.732
On right eye	2.995	1.873	1.729
On nose	3.735	2.847	2.698
On mouth	1.414	1.097	0.980
10 seconds	CG-modeled 2	CG-modeled 1	Human
On left eye	4.300	4.893	6.315
On right eye	5.544	3.385	2.529
On nose	6.676	5.622	5.812
On mouth	2.340	2.605	1.593
30 seconds	CG-modeled 2	CG-modeled 1	Human
On left eye	11.858	14.598	13.164
On right eye	14.682	12.356	8.680
On nose	18.506	16.223	14.333
On mouth	7.257	5.985	5.385

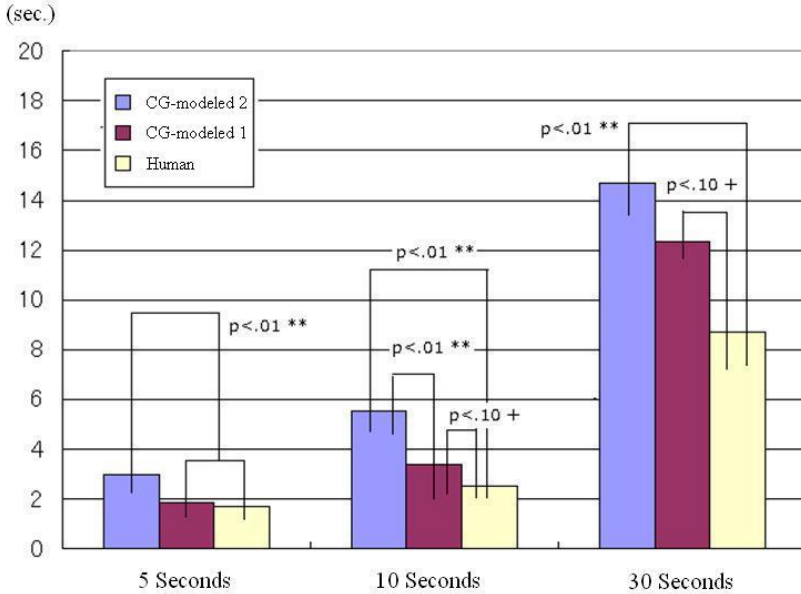


Fig. 6. Averages of time length of eye fixations on the right eyes of each face

3.3 The Questionnaires

Twelve participants (9 males and 3 females) responded to the questionnaires. Fig. 7 shows average scores of responses to Q1. The averages were significantly different between in the CG-modeled 1 and CG-modeled 2 ($p < .05$), and between in the

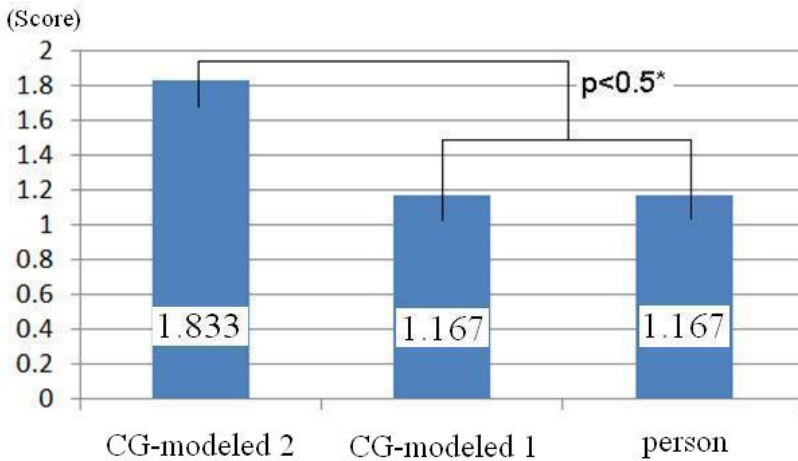


Fig. 7. Average scores of responses to Q1

CG-modeled 2 and person ($p < .05$). According to responses to Q2, the proportion of participants who described the skin and its texture was 66.7% and that of the eyes was 25.0% in the CG-modeled 1. In the CG-modeled 2, the proportion of the eyes was 50.0%, that of the mouth was 41.7% and that of the facial expression was 16.7% and the eyes was 25.0%. In the person, the proportion of the eyes was 16.7%, that of the beard was 16.7% and that of the hair was 16.7%. Moreover, the participants responded positively and negatively to CG-modeled 1 and CG-modeled 2. Most of them felt humanlike in the facial expression of CG-modeled 1 but strange in its skin and recognized easily that it was not human but CG-modeled. On the other hand, they felt humanlike in the shadows under the eyes of CG-modeled 2 but unnatural in its facial expression.

4 Discussion

4.1 Perceptual Processes in Two Steps during Judgment

The participants judged the CG-modeled 1 as non-human and the CG-modeled 2 and person as human. Time length of eye fixations on the right eye area of the CG-modeled 2 was significantly longer than that of the person in every time span. According to the responses to Q1, they found more difficult when judging the CG-modeled 2 than when the person and CG-modeled 1. Yarbus [5] mentioned that people pay much attention to unnatural or unfamiliar elements of pictures. Thus, the participants must have found the right eye of the CG-modeled 2 unnatural after observing for several seconds, and that caused their eyes to fix on it. These results are consistent with the report by Minato et al. [4].

In case of the CG-modeled 1, it was remarkable that there was no significant difference of time length of eye fixations on the right eye area from that of the person in initial 5 seconds, whereas the difference emerged as time passed. Thus, these results must imply that the participants had once perceived the CG-modeled 1 as human in the short-time observation. The emergence of the difference of time length of eye fixations must have been brought by shift of the participants' attention to differences of the CG-modeled 1 from a real human. Although the participants reported that the CG-modeled 1 was non-human, it can be assumed that people initially perceive a humanlike agent fairly similar to real human as human and then perceive it as non-human. Thus, perceptual processes of a humanlike agent can be considered to include the two steps.

4.2 The Dual-Pathway of Emotion

LeDoux [6] proposed that human brain functions and neural mechanisms of processes have two pathways of high and low roads and both of them play important roles in responding to dangerous stimuli from the external world. A stimulus is processed at first by the thalamus, and then the low road crudely leads information of the stimulus directly to the amygdala, which enables immediate response to the stimulus. On the other hand, the high road is another pathway to the amygdala via the cerebral cortex,

which simultaneously processes the stimulus carefully in detail. This dual pathway, especially response by the low road, is an essential function to avoid a danger.

The results in our experiment can be explained along with the concept of the dual pathway of emotion. When human observes a humanlike agent, the low road transmits information roughly but immediately so that he/she perceives the agent as human. Consequently, the observer initially responds to the agent as if it is human. On the other hand, the high road provides more elaborate information of the agent and that allows the observer to realize that the agent is actually non-human. These parallel functions by the dual pathway form two different perceptual processes.

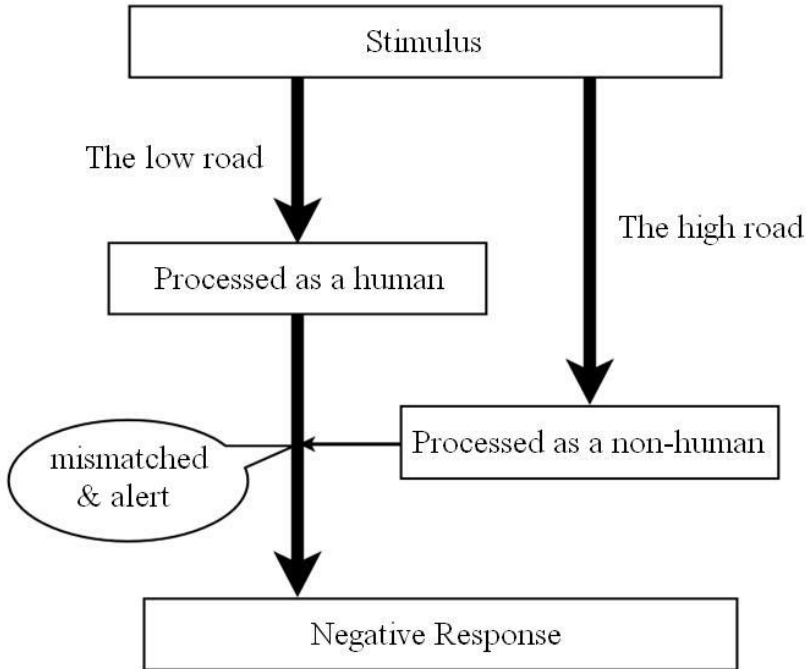


Fig. 8. Conceptual model of human perception toward a humanlike agent

This model can provide an explanation of the mental mechanism behind the uncanny valley from the viewpoint of the function to avoid dangers. For human, perceiving a humanlike agent as a human by the low road means a mistake of a different species for an individual akin to them. After the high road transmits the information, he or she receives an alert to the agent to avoid a danger which may be caused by the mistake. Accordingly, two pieces of information between the low and high road are mismatched, and the alert makes human generate negative response to the agent. Fig.8 illustrates this conceptual model.

5 Conclusion

This study experimentally investigated changes in human perceptual processes of a person and humanlike agents based on the hypothesis that differences in the perceptions emerged after observing for several seconds. The results confirm that there was no significant difference between the person and CG-modeled 2 initially, whereas the difference emerged as time passed. According to the results indicating the two steps in perceptions of a humanlike agent, we modeled the uncanny valley based on the concept of the LeDoux's dual-pathway of emotion. The model insists that mismatch between pieces of information transmitted by the low and high roads can be served as a trigger of the uncanny valley.

As described above, the model explains what causes human negative response to a humanlike agent. However, it cannot explain evocation of the feeling of eeriness and the drastic decrease of familiarity documented by the uncanny valley. One important future work is hence to clarify how the mismatch can be connected with the feeling of eeriness, and to describe the more detailed model.

References

1. Kanda, T., Miyashita, T., Osada, T., Haikawa, Y., Ishiguro, H.: Analysis of humanoid appearances in human-robot interaction. *Journal of the Robotics Society of Japan* 24(4), 497–505 (2006)
2. Mori, M.: Uncanny Valley. *Energy* 7(4), 33–35 (1970)
3. Noma, M., Saiwaki, N., Itakura, S., Ishiguro, H.: Composition and Evaluation of the Humanlike Motions of an Android. In: *Proc. Int. Conf. Humanoid Robots*, pp. 163–168 (2006)
4. Minato, T., Shimada, M., Ishiguro, H., Itakura, S.: Development of an Android Robot for Studying Human-Robot Interaction. In: *Proc. IEA/AIE. Conf. 2004*, pp. 424–434 (2004)
5. Yabus, A.L.: *Eye Movements and Vision*. Preum Press, New York (1967)
6. LeDoux, J.E.: *The Emotional Brain: The Mysterious Underpinnings of Emotional Life*. Simon & Schuster (1998)

Multi-party Human-Machine Interaction Using a Smart Multimodal Digital Signage

Tony Tung, Randy Gomez, Tatsuya Kawahara, and Takashi Matsuyama

Kyoto University,
Academic Center for Computing and Media Studies
and Graduate School of Informatics, Japan
tung@vision.kuee.kyoto-u.ac.jp,
{randy-g,kawahara}@ar.media.kyoto-u.ac.jp, tm@i.kyoto-u.ac.jp

Abstract. In this paper, we present a novel multimodal system designed for smooth multi-party human-machine interaction. HCI for multiple users is challenging because simultaneous actions and reactions have to be consistent. Here, the proposed system consists of a digital signage or large display equipped with multiple sensing devices: a 19-channel microphone array, 6 HD video cameras (3 are placed on top and 3 on the bottom of the display), and two depth sensors. The display can show various contents, similar to a poster presentation, or multiple windows (e.g., web browsers, photos, etc.). On the other hand, multiple users positioned in front of the panel can freely interact using voice or gesture while looking at the displayed contents, without wearing any particular device (such as motion capture sensors or head mounted devices). Acoustic and visual information processing are performed jointly using state-of-the-art techniques to obtain individual speech and gaze direction. Hence displayed contents can be adapted to users' interests.

Keywords: multi-party, human-machine interaction, digital signage, multimodal system.

1 Introduction

Over the last decade smart systems for multi-party human-machine interaction have become ubiquitous in many everyday life activities (e.g., digital advertising displays, video games or poster presentations). Here, we present a novel multimodal system that is designed for smooth multi-party human-machine interaction. The system detects and recognizes verbal and non-verbal communication signals from multiple users, and returns feedbacks via a display screen. In particular, visual information processing is used to detect communication events that are synchronized with acoustic information (e.g., head turning and speech). We believe the system can potentially be adapted to various applications such as entertainment (multiplayer interactive gaming device), education or edutainment (virtual support for lecturer), medicine, etc.

Multimodal Audio/Video systems designed for human behavior and interaction analysis usually consist of multiple video cameras and microphones placed

in a dedicated room, and oriented towards participants. To date, these systems are still very tedious to setup and often require wearable equipments that prevent them to be used casually or in an uncontrolled environment. Here, the proposed system is portable and scalable. It consists of a digital signage or large display equipped with multiple sensing devices spaced on a portable structure: a microphone array, 6 HD video cameras, and two depth sensors. The display is used to show various contents, such as poster presentations, web browsers, photos, etc. On the other hand, multiple users standing in front of the panel can interact freely using voice or gesture while looking at the displayed contents, without wearing any particular device (such as motion capture sensors or head mounted devices). We tested the system with real poster presentations as well as casual discussions. The next sections present related work, description of the framework, A/V multimodal data processing, application to multi-party interaction, and conclusion.

2 Related Work

Human-to-human and human-computer interaction have been studied in numerous fields of science such as psychology, computer graphics, communication, etc. In group communication, humans use visual and audio cues to convey and exchange information. Hence video and audio data have been naturally extensively used to study human behavior in communication. For example, several corpus such as VACE [Chen et al., 2006], AMI [Poel et al., 2008], Mission Survival [Pianesi et al., 2007], IMADE [Sumi et al., 2010] were created to capture multimodal signals in multi-party conversation and interaction. Speech is often used to detect dominant speakers based on turn-taking and behavior analysis, while non-verbal cues provide feedbacks to understand communication patterns and behavior subtleties (e.g., smile, head nodding or gaze direction change) and can be used as back-channels to improve communication [White et al., 1989]. Nevertheless heavy equipments (e.g., headsets) are often required, and visual information processing is usually limited (due to video resolution). Other systems using digital signage (like the moodmeter from MIT) usually use only one video camera that performs only face detection/classification. Acoustic is not used and they do not consider human-machine interaction. Commercial systems, like Samsung Smart TVs, use single-human gestures as remote control and do not handle interaction with multiple people. To our knowledge, no framework has been proposed in the literature that aims at multi-people interacting with a digital signage using multimodal (audio and video) signals. Note that in [Tung et al., 2012], the authors introduce a multimodal interaction model for multi-people interaction analysis that can be used with the system presented in this paper. As well, let us cite [Tung and Matsuyama, 2012] in which the authors present 3D scene understanding using data from multiview video capture.

3 Smart Multimodal System Configuration

Audio. Microphone array is employed as audio capturing device. The spatial arrangement of the microphone sensors enable the system to localize different

sound sources: the speech signal that originates from the individual participants and the signals that come from other sources (i.e., background noise). The microphone array system provides a hands-free audio capture mechanism in which the participants are not constrained to using wired microphones that limits movements, paving the way towards free flowing interaction. The power of the captured audio signals from the distant sources are improved by increasing the number of microphone sensors in the array, in this case, a reliable level of speech power is achieved which is very important in the speech recognition system. The microphone array is configured linearly with 19 microphone sensors and placed on top of a 65-inch display display (see Fig. 1) and the sampling rate is 16KHz. The smart audio processing in a poster presentation scenario (i.e., participants: presenter and audience) is described in Sec. 4.

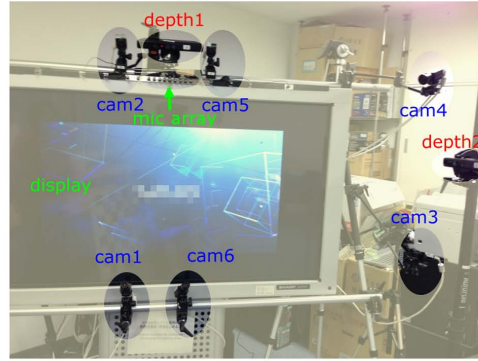
Video. Multiple video cameras are employed to capture nonverbal communication and interaction between multiple people. In order to capture multiple views of subjects standing in front of the system, 6 HD video vision cameras (from Point Grey) are spaced on a portable structure made of poles and mounted around a display. In the context of poster presentation, sensing devices are placed at on one side of the display to capture a presenter, and at the center to capture the audience (e.g., two or three people). We place 3 cameras with 3.5mm lenses on top of the display (two at the center, one on the side) to obtain wide field-of-view (150 deg), perform video capture in UXGA at 30 fps, and 3D reconstruction from stereo. As well, 3 cameras with 12mm lenses are placed below the display (two at the center, one on the side) to capture closeup videos in SVGA of users' faces at 30 fps. As well, two depth sensors (MS Kinect) are placed on top of the screen (one at the center, one on the side) and capture depth map videos in VGA at 30 fps. To make the system easily transportable, only one PC with a single GPU is used for video capture and processing. Fig. 1 shows the current system and multiview video samples.

4 Multimodal Data Processing

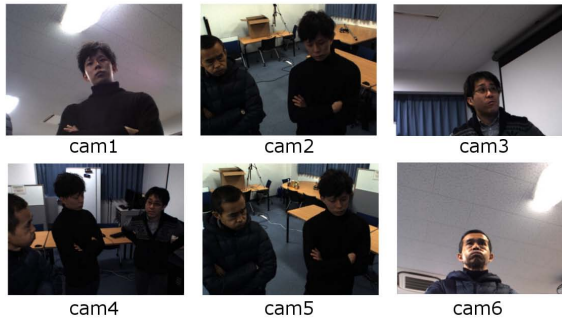
Audio. Acoustic signal processing in a multi-party system involves the processing of the data captured from the microphone array and the design of the automatic speech recognition (ASR) system. The multimodal signage application can be used in various scenarios, specifically in this paper we focus on poster presentation that involves the interaction between the presenter and the audience. Audio processing is described as follows:

- Microphone Array Processing

Assuming that there are N sources (i.e., coming from participants) and M ($\geq N$) microphone sensors (in our case $M=19$). Let us denote $\mathbf{s}(\omega)$ as the input acoustic signal of N sources in frequency domain, described as $\mathbf{s}(\omega) = [s_1(\omega), \dots, s_N(\omega)]^T$, where T represents the transpose operator. The received signals in vector form is denoted as $\mathbf{o}(\omega) = [o_1(\omega), \dots, o_M(\omega)]^T$.



a) Multimodal system



b) Multiview video

Fig. 1. a) Multimodal system setup for smart poster presentation. b) Multiview video capture of one presenter and two attendees.

Microphone array signal processing is described as follows:

$$\mathbf{o}(\omega) = \mathbf{A}(\omega)\mathbf{s}(\omega) + \mathbf{n}(\omega), \quad (1)$$

where $\mathbf{A}(\omega) \in \mathbb{C}^{M \times N}$ is the *Room Impulse Response (RIR)* in matrix form. The RIR describes the room characteristics that governs the behaviour of the sound signal as it is reflected inside the room enclosure. The background noise is denoted by $\mathbf{n}(\omega)$. We note that the both $\mathbf{n}(\omega)$ and $\mathbf{s}(\omega)$ are assumed to be statistically independent which is usually true in real environment conditions dealing with simple types of noise contamination. The sound sources are spatially separated using *Geometrically constrained High-order Decorrelation based Source Separation (GHDSS)*, which is a byproduct of both beam forming and blind separation [Sawada et al., 2002][Nakajima et al., 2008]. The separated signal is denoted as $\hat{s}^{(l)}(\omega)$.

– Context and Models for ASR

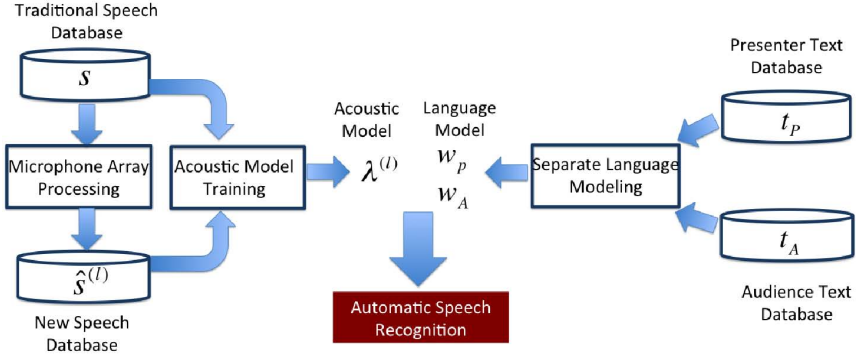


Fig. 2. Robust modeling for automatic speech recognition system

We trained separate acoustic models based on different user profiles. During the early stage of the design process, roles of the participants are profiled in advance. For example, if the system is used for poster presentation, participants are classified as presenter, audience, etc. The task of the participants are known right from the very beginning; the presenter usually has a script which is used to discuss the content of the presentation displayed on the digital signage. In most cases, the signage itself contains information of the presenters talk (i.e., text). Moreover, the audience is presumed to ask questions, and because of this prior information, it is safe to assume that the conversation dynamics between the participants are readily available. Consequently, depending on the directivity of the speaker we can re-train the acoustic model to $\lambda^{(l)}$ with data processed using the sound separation mechanism discussed above. This will improve performance as opposed to solely using the baseline model λ trained from the traditional speech database because the actual location (l) of the speakers are considered in the former. Separate language models can also be modeled for the presenter and the audience, respectively. By using the corresponding text data that are unique to the conversation dynamics to each of the participant class, separate language models are trained: w_p (for presenter) and w_a (for audience). Training procedure that involves context is shown in Fig. 2. In our method, we are able to break down the generic approach in the automatic speech recognition system (ASR). By incorporating context with respect to the microphone array processing reflective of the participant's position and the separate language models, the system can automatically switch parameters that are appropriate to the current scenario. In the end, robustness is achieved.

Video. Visual information processing is achieved using the combination of depth sensors that deliver depth maps of the scene in VGA resolution, and multiple video cameras that capture the scene in higher resolution (UXGA and SVGA). All multiview video cameras are geometrically calibrated using standard methods (i.e., with a chessboard) and synchronized by software trigger.

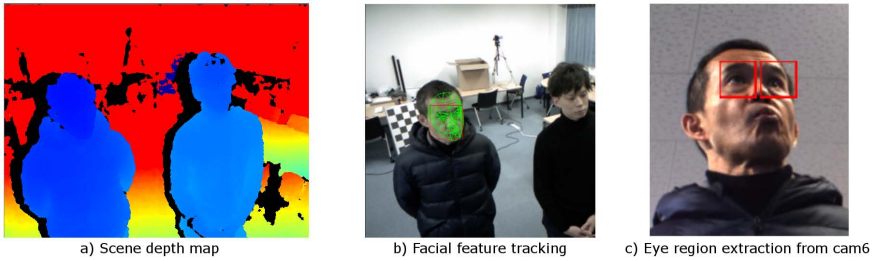


Fig. 3. a) Depth map from frontal depth sensor. b) Face feature tracking using depth and color information (by Kinect SDK). c) Eye localization after registration to HD video camera 6.

The depth sensors (IR cameras) are also calibrated with the HD video cameras. Depth data serve for human body-part identification using approaches based on discriminative random regression forests that efficiently label different body regions [Shotton et al., 2011]. As well, head pose and face feature tracking can be performed from depth and color information of the depth sensors using state-of-the-art techniques [Cootes et al., 2006][Viola et al., 2001][Fanelli et al., 2011]. As the resolution of color cameras integrated in current consumer depth sensors are usually too poor to provide accurate gaze estimation, HD video cameras placed below the screen are used instead for accurate gaze direction estimation [Xu et al., 2008][Feng et al., 2011]. In practice, as shown in Fig. 3 we extract regions of interest (e.g., faces and eyes), and register HD video frames to the depth maps. Compared to prior multimodal setup [Tung et al., 2012], the proposed design is able to provide more accurate head pose and gaze estimation based on eye detection. Note that, as in [Tung et al., 2012] HD video cameras placed on top of the screen provide rough depth maps in real-time (using multiview stereo reconstruction), which can be merged with data from the depth sensors for further depth accuracy enhancement. Furthermore, head and mouth positions are used in the speech processing described above for better diarization.

5 Applications: Multi-party Multimodal Interaction

In this work, we combine acoustic and video information for seamless interaction with smart display. Suppose that a display contains text data localized into several regions according to visual coordinate locations. The generic word vocabulary which is composed of the total text is broken down into sub-regions, corresponding to the several visual coordinate locations. Using the localized vocabulary (within a particular region), a new set of language model is updated for each region. This allows us to dynamically select active regions on the display based on the gazes of the participants. The system dynamically switches language models for speech recognition, reflective of the active region. This strategy minimizes the confusion in speech recognition when displaying multiple contents,



Fig. 4. When multiple browsers are opened simultaneously, it is not trivial for hands-free voice systems to understand what applications users are focusing on. The proposed system uses gaze and speech processing to determine what actions to perform.

as a particular word entry may occur several times in different locations within the whole display. For example, as illustrated in Fig. 4 when multiple browsers are opened simultaneously, it is not trivial for hands-free voice systems to understand what applications users are thinking about when simple actions are requested (e.g., close browser, search in window, etc.).

Switching to active regions based on the users' gaze narrows down the size of the vocabulary as defined by each active region. In effect, the system benefits from both acoustic and visual information in improving overall system performance. Furthermore, the system can be used as a smart poster for automatic presentation to a group of people. By capturing and analyzing gaze directions and durations, attention or interest levels can be derived for each portion of the presentation (see [Tung et al., 2012]). Hence, when joint-attention is detected (e.g., when two subjects are looking at the same region), the system can automatically highlight the region of interest. We believe the proposed system can be used for numerous applications, such as education, medicine, entertainment, and so on.

6 Conclusion

We present a novel multimodal system that is designed for smooth multi-party human-machine interaction. The system detects and recognizes verbal and non-verbal communication signals from multiple users, and returns feedbacks via a display screen. In particular, visual information processing is used to detect communication events that are synchronized with acoustic information (e.g., head turning and speech). To our knowledge, no similar setup has been proposed yet in the literature.

Acknowledgment. This work was supported in part by the JST-CREST project Creation of Human-Harmonized Information Technology for Convivial

Society. The authors would like to thank Hiromasa Yoshimoto for his work on system development and data capture.

References

- [Chen et al., 2006] Chen, L., Rose, R.T., Qiao, Y., Kimbara, I., Parrill, F., Welji, H., Han, T.X., Tu, J., Huang, Z., Harper, M.P., Quek, F., Xiong, Y., McNeill, D., Tuttle, R., Huang, T.: Vace multimodal meeting corpus. In: Renals, S., Bengio, S. (eds.) MLMI 2005. LNCS, vol. 3869, pp. 40–51. Springer, Heidelberg (2006)
- [Cootes et al., 2006] 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)
- [Fanelli et al., 2011] Fanelli, G., Weise, T., Gall, J., Van Gool, L.: Real Time Head Pose Estimation from Consumer Depth Cameras. In: Mester, R., Felsberg, M. (eds.) DAGM 2011. LNCS, vol. 6835, pp. 101–110. Springer, Heidelberg (2011)
- [Feng et al., 2011] Feng, L., Sugano, Y., Okabe, T., Sato, Y.: Inferring human gaze from appearance via adaptive linear regression. In: ICCV (2011)
- [Gomez and Kawahara, 2010] Gomez, R., Kawahara, T.: Robust speech recognition based on dereverberation parameter optimization using acoustic model likelihood. *IEEE Trans. Audio, Speech and Language Processing* 18(7), 1708–1716 (2010)
- [Nakajima et al., 2008] Nakajima, H., Nakadai, K., Hasegawa, Y., Tsujino, H.: Adaptive Step-size Parameter Control for real World Blind Source Separation. In: ICASSP (2008)
- [Pianesi et al., 2007] Pianesi, F., Zancanaro, M., Lepri, B., Cappelletti, A.: A multimodal annotated corpus of concensus decision making meetings. *Language Resources and Evaluation*, 409–429 (2007)
- [Poel et al., 2008] Poel, M., Poppe, R., Nijholt, A.: Meeting behavior detection in smart environments: Nonverbal cues that help to obtain natural interaction. In: FG (2008)
- [Sawada et al., 2002] Sawada, H., Mukai, R., Araki, S., Makino, S.: Polar coordinate based nonlinear function for frequency-domain blind source separation. In: ICASSP (2002)
- [Shotton et al., 2011] Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-Time Human Pose Recognition in Parts from a Single Depth Image. In: CVPR (2011)
- [Sumi et al., 2010] Sumi, Y., Yano, M., Nishida, T.: Analysis environment of conversational structure with nonverbal multimodal data. In: ICMI-MLMI (2010)
- [Tung and Matsuyama, 2012] Tung, T., Matsuyama, T.: Topology Dictionary for 3D Video Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 34(8), 1645–1657 (2012)
- [Tung et al., 2012] Tung, T., Gomez, R., Kawahara, T., Matsuyama, T.: Group Dynamics and Multimodal Interaction Modeling using a Smart Digital Signage. In: Fusiello, A., Murino, V., Cucchiara, R. (eds.) ECCV 2012 Ws/Demos, Part I. LNCS, vol. 7583, pp. 362–371. Springer, Heidelberg (2012)
- [Viola et al., 2001] Viola, P., Jones, M.: Robust real-time object detection. In: IJCV (2001)
- [White et al., 1989] White, S.: Backchannels across cultures: A study of americans and japanese. *Language in Society* 18, 59–76 (1989)
- [Xu et al., 2008] Xu, S., Jiang, H., Lau, F.C.: User-oriented document summarization through vision-based eye-tracking. In: 13th ACM Int'l Conf. Intelligent User Interfaces (2008)

A Remote Pointing Technique Using Pull-out

Takuto Yoshikawa, Yuusaku Mita, Takuro Kuribara,
Buntarou Shizuki, and Jiro Tanaka

University of Tsukuba, Japan
{yoshikawa,mita,kuribara,shizuki,jiro}@iplab.cs.tsukuba.ac.jp

Abstract. Reaching objects displayed on the opposite side of a large multi-touch tabletop with hands is difficult. This forces users to move around the tabletop. We present a remote pointing technique we call *HandyPointing*. This technique uses pull-out, a bimanual multi-touch gesture. The gesture allows users to both translate the cursor position and change control-display (C-D) ratio dynamically. We conducted one experiment to measure the quantitative performance of our technique, and another to study how users selectively use the technique and touch input (i.e., tap and drag).

Keywords: bimanual interaction, multi-touch, gesture, tabletop.

1 Introduction

A large multi-touch tabletop is used for a collocated collaborative work that involved multiple users. These users surround the tabletop and touch the tabletop from their respective positions. However, reaching a distant object displayed on the opposite side of the tabletop is difficult due to the largeness of the touchscreen. Toney et al. reported that more than 90% of users' touch interactions are performed within 34 cm from their respective positions [11]. Users are forced to lean forward from the tabletop or move around the tabletop to reach the object.

To solve this problem, indirect-pointing devices, such as a mouse, are complementary used for touch input. Using these devices enable users to reach the distant objects. However, they require physical space in which to place them around a tabletop for each user. Furthermore, preparing the devices in advance is troublesome because tabletops are used by an unspecified number of users simultaneously.

Therefore, we present a remote pointing technique we call *HandyPointing*. This technique uses pull-out, a bimanual multi-touch gesture [12], to determine a cursor position. A pull-out gesture requires no additional devices because the technique uses touch input only. Furthermore, a pull-out gesture allows users not only to translate the cursor position but also to change the control-display (C-D) ratio dynamically, similar to [3]. This means that they can selectively perform rough pointing with a large C-D ratio and precise pointing with a small C-D ratio. Therefore, users can precisely point at a distant position quickly by combining these rough and precise pointing techniques.

2 Related Work

There are related works about remote pointing techniques on tabletops. We classify these techniques into direct-pointing, and indirect-pointing. Here the former uses the position at which users point as a cursor position, and the latter translates a cursor position according to the movements of devices.

Direct-pointing. Parker et al. used the shadow of the tip of stylus to point at a distant position [9]. In the work of Banerjee et al. [3], users could point a finger at objects on tabletops and dynamically change C-D ratio by one hand while performing a pointing with the other hand. The above techniques required additional devices that obtain the position of users' hands to realize direct-pointing. In HandyPointing, we adopt indirect-pointing in order not to use such devices. Our technique requires only the touch coordinates.

Indirect-pointing. Bartindale et al. [4] and Matejka et al. [8] developed an onscreen mouse for multi-touch tabletops that allows users to point, similar to a conventional physical mouse. These research realized an indirect-pointing technique. However, they required to recognize the shape of hands, while our technique can be applied to tabletops that detect more than three touch points. I-Grabber [1] is an onscreen widget manipulated by multi-touch interactions. Users can select and translate a distant object with the widget. Although our technique also uses multi-touch interactions, the technique allows users to determine a cursor position by a single multi-touch gesture.

Bimanual Interaction. Guiard modeled the asymmetric bimanual behaviors of humans as Kinetic Chain Model [6]. Tokoro et al. presented a pointing technique that utilized two acceleration sensors, and postures of both hands determined a cursor position [10]. Furthermore, Malik et al. developed a bimanual pointing technique by using image processing [7]. In contrast with these techniques, our technique is realized by using touch based gestures. In addition, Bailly et al. utilized the number of touches and their strokes of both hands to execute commands in a distant menu bar [2]. While their technique enables command invocations that utilized discrete input, our technique enables indirect-pointing that utilized continuous input by the stroke of a pull-out gesture.

3 Interaction Techniques

Our pointing technique utilizes both hands to move a cursor. In this section, we describe HandyPointing, a remote pointing technique, and its additional remote manipulation technique.

3.1 Pointing Technique

Figure 1 shows the procedure of HandyPointing. First, users put two fingers of their non-dominant hand (base-fingers) on a tabletop, as shown in Figure 1a.

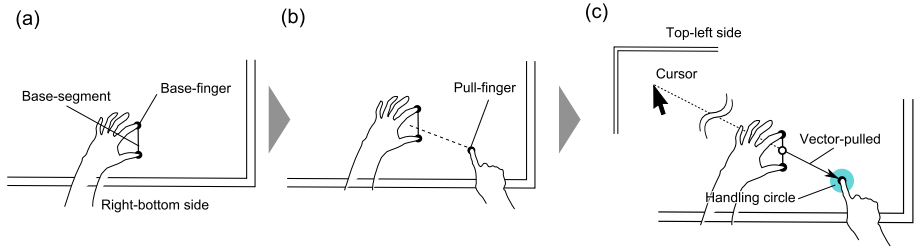


Fig. 1. Pointing procedure using HandyPointing

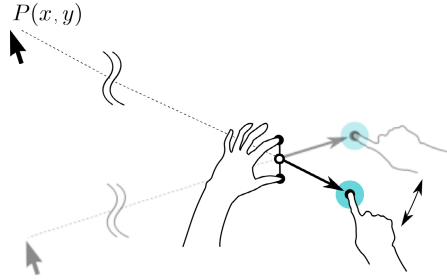


Fig. 2. Cursor translation according to vector-pulled

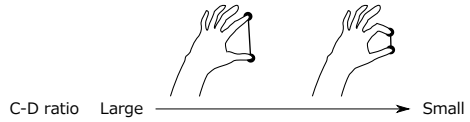


Fig. 3. Dynamic C-D ratio according to length of base-segment

When users drag their finger of their dominant hand (pull-finger) to cross the segment between base-fingers (base-segment) as shown in Figure 1b, a cursor is displayed on an extension of the opposite direction of the vector from the center of the base-segment to pull-finger (vector-pulled) as shown in Figure 1c. Users can quit pointing by taking base-fingers off from the tabletop.

If users arrange the vector-pulled, the cursor position changes in accordance with the vector as shown in Figure 2. C-D ratio also changes depending on the length of the base-segment. This means that users can simultaneously move the cursor by using the dominant hand while controlling C-D ratio dynamically by using the other hand as shown in Figure 3.

An advantage of this bimanual manipulation is that users can selectively perform rough pointing with a large C-D ratio or precise pointing with a small C-D ratio. For example, users can move a cursor precisely with a short base-segment, while they can quickly move the cursor at a distance with a long base-segment as shown in Figure 4.

3.2 Determination of a Cursor Position

This section describes the procedure to determine a cursor position. Suppose that $P_i(x, y)$ is the i -th cursor position after i frames have passed since users placed base-fingers on the tabletop. Then P_i is given by the following expressions:

$$P_i = G_0 - \sum_j^i k_j \Delta V_j,$$

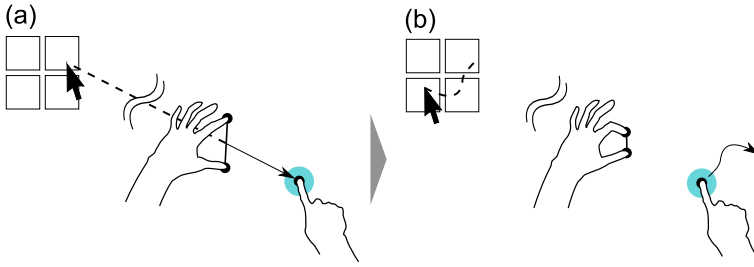


Fig. 4. Usage of dynamic C-D ratio. Users (a) point at far position quickly with large C-D ratio, and then (b) precisely point at object with small C-D ratio.

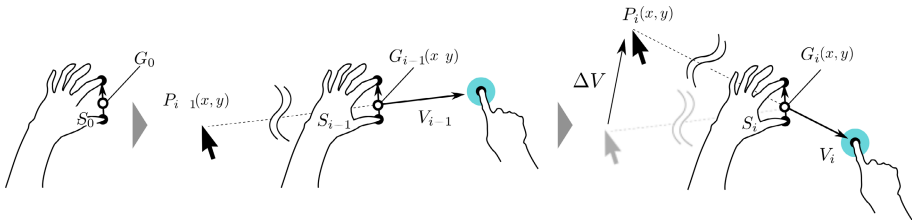


Fig. 5. Determination of cursor position

$$\Delta V_i = V_i - V_{i-1},$$

$$k_i = \alpha \times \frac{|S_0|}{|S_i|}.$$

As shown in Figure 5, S_0 is the base-segment when base-fingers were placed on the tabletop, and G_0 is the gravity point of S_0 . Furthermore, S_i and V_i are the i -th base-segment and vector-pulled, respectively. Then, G_i is the gravity point of S_i . α is a constant. That is, our technique determines i -th C-D ratio by k_i , and then the cursor position P_i is moved by k_i and ΔV_i , which is the difference of V_i , frame by frame.

3.3 Remote Manipulation

We implemented a function to manipulate a distant object. As shown in Figure 1c, a circle is shown around the pull-finger (handling-circle) when users begin pull-out. Users can select the object under the cursor by tapping the handling-circle after they have translated a cursor. Moreover, they can translate the cursor again by dragging the handling-circle. They can unselect the object by tapping the circle again.

The selected object moves according to the cursor when users drag the handling-circle. This means that they can select a distant object, and drag it to another distant location. Note that they can dynamically change C-D ratio while they are using not only a normal pointing but also a remote manipulation. Therefore, they can select the object precisely and move it quickly.

4 Evaluation

We conducted two experiments. One was to measure the quantitative performance of HandyPointing, and the other was to study how users selectively use HandyPointing and touch input (i.e., tap and drag). We implemented a prototype of HandyPointing using a 1470mm x 800mm 60-inch tabletop. Ten volunteers participated in this experiment. They were undergraduates or graduate students with ages ranging in age from 21 to 24 years. They were all right-handed and familiar with a mouse.

4.1 Experiment 1

To compare the performance of HandyPointing and another in-direct pointing technique, we used a pointing task similar to those in [3, 5]. We used a mouse as it is the most common in-direct pointing device. We divided participants into two groups. One group first performed the task by HandyPointing; the other group first performed the task by mouse. We asked participants to sit in a chair that placed in the middle of the short side of the screen during the task. We explained how to use each technique before the task, and asked them to practice the techniques sufficiently. To measure the base-line of our technique, we assigned the C-D ratio of the mouse to achieve the best performance such that the mouse never needed to be clutched.

Task. We asked participants to point at and select a target object. When participants selected a start point, a target object was displayed. We radially arranged the positions of target objects as shown in Figure 6. Once participants selected the target object, the object was removed.

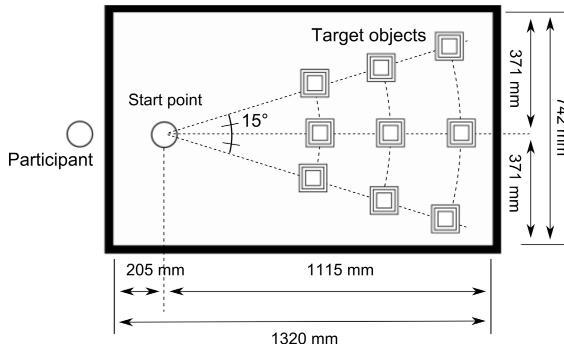


Fig. 6. Positions of start point and target objects

In this experiment, independent variables were: target distance (500, 700, and 900 pixels, i.e. approximately 515mm, 715mm, and 915mm, respectively), target angle (-15, 0, and 15 degree), target size (40, 80, and 120 pixels, i.e. approximately 41mm, 82mm, and 123mm, respectively), and technique (HandyPointing

and mouse). Each participant performed 2 trials for each combination of factors, thus they performed 108 (3x3x3x2x2) trials in total. All trials were presented in a randomized order.

Result. We measured the time to complete a trial (trial-time) and the number of errors. In HandyPointing condition, trial-time begins when participants started HandyPointing on a start point and ends when they selected a target object by tapping a handling-circle. In mouse conditions, trial-time begins when participants clicked a start point and ends when they selected a target object. When participants failed to select a target object, we treated it as an error.

Figure 7 to 10 show the result of the experiment. Figure 7 and 8 show the average trial-time and the number of errors in HandyPointing condition. Figure 9 and 10 show those in mouse condition. In these figures, the blue graph illustrates the result for each target distance (500, 700, and 900 pixels), green one illustrates the result for each target angle (-15, 0, and 15 degree) and red one illustrates the result for each target size (40, 80, and 120 pixels). The average trial-times were 3812ms in HandyPointing condition and was 1266ms in mouse condition. The average numbers of errors were 1.72 in HandyPointing condition and was 0.053 in mouse condition. Figure 11 and 12 show average trial-time of HandyPointing trials and mouse trials for each participant.

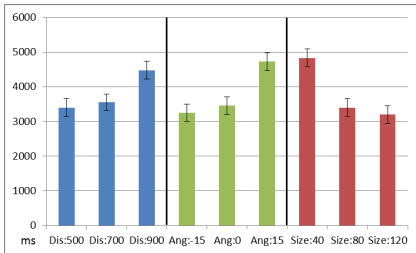


Fig. 7. Mean and variance of trial-time for each target distance (blue), target angle (green), and target size (red) in HandyPointing condition

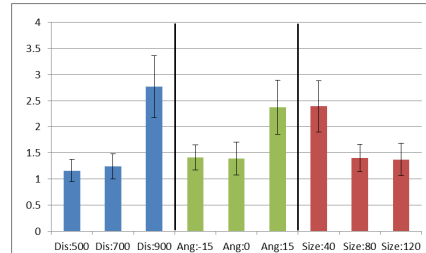


Fig. 8. Mean and variance of number of errors for each target distance (blue), target angle (green), and target size (red) in HandyPointing condition

Discussion. The trial-time of HandyPointing was larger than that of a mouse, as shown in Figure 7 and 9. However, trial-time gradient and the number of errors were similar for each condition. That is, trial-time was in accordance with the increase in distance and the decrease in angle. This result indicates that HandyPointing seems to follow Fitts' law.

The number of errors of HandyPointing increased when target distance was 900 pixels, as shown in Figure 7. This means participants failed to tap a handling-circle when they selected the most distant targets. This is because they could not tap the circle while focusing on a distant target object. On the other hand, participants

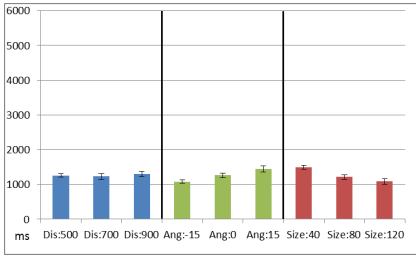


Fig. 9. Mean and variance of trial-time for each target distance (blue), target angle (green), and target size (red) in mouse condition

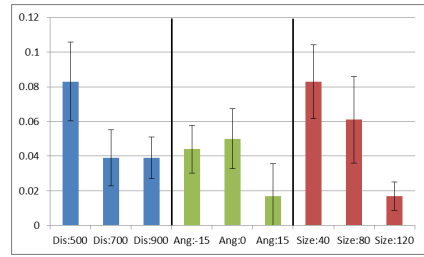


Fig. 10. Mean and variance of number of errors for each target distance (blue), target angle (green), and target size (red) in mouse condition. Note that maximum of Y-axis is different from Figure 8.

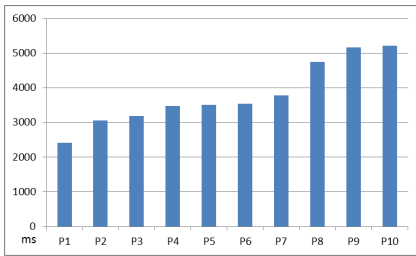


Fig. 11. Mean of trial-time for each participant in HandyPointing condition

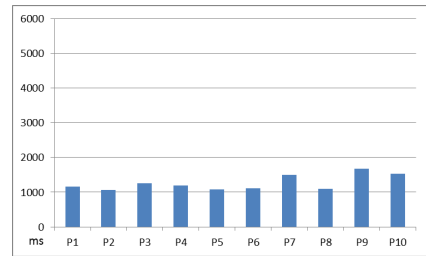


Fig. 12. Mean of trial-time for each participant in mouse condition

successfully tapped the circle more when they selected a near target object. This is because they could focus on the circle and the object simultaneously.

Furthermore, the trial-time and the number of errors also increased in HandyPointing condition as the angle changed from -15 to 15 degree, as shown in Figure 7 and 8. This increase was more significant than for the mouse. A cursor is displayed on the non-dominant hand side first, when they begin to perform HandyPointing. For example, the cursor is displayed on the left side for right-handed participants. Thus, they can easily select the non-dominant hand side objects. However, selecting the dominant hand side objects forces participants to move arms outside of the natural range of motion of arms. Hence, they had to move their right arms to the left side and left arms to the right side. This makes it difficult to select the dominant hand side objects and shows that users can easily select non-dominant hand side objects by using HandyPointing.

Next, we discuss average trial-time for each participant. These trial-times are shown in Figure 11 and 12. In HandyPointing condition, the largest average trial-time (5211ms) was 2.15 times larger than the smallest average trial-time (2420ms). In mouse condition, the largest average trial-time (1524ms) was 1.57 times larger than the smallest average trial-time (1165ms). Moreover, the

smallest average trial-time of HandyPointing (2420ms) was 2.08 times larger than that of a mouse (1165ms). Trial-time greatly differed among participants. In this experiment, we asked participants to practice HandyPointing and a mouse, until they had become familiar with these techniques. However, some participants proceeded the experiment as soon as they had learned only how to use HandyPointing but without becoming familiar with the technique. This meant the degree of proficiency in HandyPointing for each participant was significantly different, resulting in the large differences in trial-time among participants. On the other hand, they were already familiar with the mouse. Thus, the differences in trial-time in mouse condition was small. This result means that the degree of proficiency significantly affects the performance of HandyPointing among participants. However, with a little training, pointing with HandyPointing takes at least only twice time as long as pointing with a mouse. In addition, more training may lead to better performance.

4.2 Experiment 2

To study how users selectively use HandyPointing and ordinary touch input, we used select & docking task, similar to those in [3, 5]. This experiment was sequentially conducted after Experiment 1. This means that the same participants joined this experiment, and the same apparatuses were used.

Task. The same as Experiment 1, we asked participants to point at an object, select it, and move it into a dock. In this task, objects appeared on the top, middle, or bottom of the screen as illustrated in Figure 13, and docks appeared near or far from participants. Participants were allowed to use touch input and HandyPointing simultaneously during the experiment.

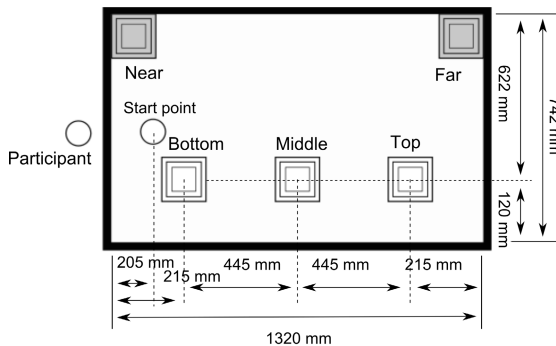


Fig. 13. Positions of target objects and docks

In this experiment, independent variables were: target size (40, 80, and 120 pixels, approximately 40mm, 80mm, and 120mm respectively), target position (top, middle, and bottom), and dock position (near and far). The dock was the same size as the target in each task. Each participants performed 3 trials for

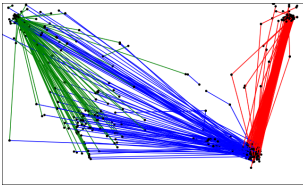


Fig. 14. Movements of top objects

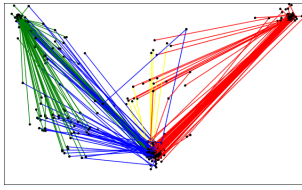


Fig. 15. Movements of middle objects

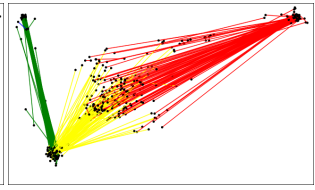


Fig. 16. Movements of bottom objects

each combination of factors, thus they performed 54 ($3 \times 3 \times 2 \times 3$) trials in total. All trials were presented in a randomized order.

Result. Figure 14 to 16 shows the movements of objects, where blue and green segments represent the movements of objects toward a near dock that are manipulated by HandyPointing and touch input, respectively. Red and yellow ones represent those of objects toward a near dock that are manipulated by HandyPointing and touch input, respectively. The segments connect the positions where objects began and the stopped moving.

Figure 14 shows the movements of top objects. All top objects toward a near dock were manipulated by HandyPointing only, and no manipulations were performed by touch input. In contrast, those toward a near dock were manipulated by the combination of HandyPointing and touch input. This means participants used HandyPointing to move a distant object to the near position and touch input to move it into the near dock. Figure 15 shows the movements of middle objects. 89% of the objects toward a near dock were manipulated by HandyPointing only, and the remaining 11% were done by the combination of HandyPointing and touch input. This is because some participants first moved the objects to the near position by touch input and then into the far dock by HandyPointing. 39% of the middle objects toward a near dock were manipulated by touch input only, and 3% were performed by HandyPointing only. The remaining 58% were performed by the combination of HandyPointing and touch input. Figure 16 shows the movements of bottom objects. All objects toward a near dock were manipulated by the combination of HandyPointing and touch input, and 99% of the objects toward a near dock were manipulated by touch input.

Discussion. Participants' behavior changed depending on the distance to a target object. Participants used HandyPointing to drag top and middle objects into a near dock. They used the combination of HandyPointing and touch input to drag them into a near dock (Figure 14 and 15). In contrast, they used the combination of HandyPointing and touch input to drag bottom objects into a near dock and used touch input to drag them into a near dock (Figure 16). This difference is owing to whether they can reach the objects with hands. That is, participants used HandyPointing for a distant object and touch input for a near

object. The result shows that they selectively used one of these techniques in accordance with the distance to a target object.

When they combined the techniques, participants first put an object into the near position by using one technique, and then they dragged it into a dock by using the other technique. In our observation, they seemed to use the first technique to move the object into the position where they could select easily with the second technique. This indicates that they preferred to use HandyPointing for a distant object and touch input for a reachable object.

5 Conclusions and Future Work

We designed and implemented a remote pointing technique, HandyPointing. The technique allows users to point at a distant position that their hands cannot reach. Furthermore, users can simultaneously change C-D ratio dynamically by using their non-dominant hand while determining a cursor position by using the dominant hand. Therefore, they can selectively use rough pointing or precise pointing. We conducted two experiments. Their results showed that users can selectively use HandyPointing and ordinary touch input, and pointing with HandyPointing at least takes only about twice as long as pointing with a mouse with a little training.

We will continue to measure the performance of HandyPointing because the C-D ratio of a mouse in the experiments was the ratio that maximizes the performance of the mouse to measure the base-line of our technique. Therefore, we will conduct a similar experiment to measure the maximum performance of our technique by seeking the ideal C-D ration in the future and compare it with the results in this paper.

References

1. Abednego, M., Lee, J.H., Moon, W., Park, J.H.: I-Grabber: expanding physical reach in a large-display tabletop environment through the use of a virtual grabber. In: Proc. of ITS 2009, pp. 61–64 (2009)
2. Bailly, G., Lecolinet, E., Guiard, Y.: Finger-count & radial-stroke shortcuts: 2 techniques for augmenting linear menus on multi-touch surfaces. In: Proc. of CHI 2010, pp. 591–594 (2010)
3. Banerjee, A., Burstyn, J., Girouard, A., Vertegaal, R.: Pointable: an in-air pointing technique to manipulate out-of-reach targets on tabletops. In: Proc. of ITS 2011, pp. 11–20 (2011)
4. Bartindale, T., Harrison, C., Olivier, P., Hudson, S.E.: SurfaceMouse: supplementing multi-touch interaction with a virtual mouse. In: Proc. of TEI 2011, pp. 293–296 (2011)
5. Forlines, C., Wigdor, D., Shen, C., Balakrishnan, R.: Direct-touch vs. mouse input for tabletop displays. In: Proc. of CHI 2007, pp. 647–656 (2007)
6. Guiard, Y.: Asymmetric division of labor in human skilled bimanual action: The kinematic chain as a model. *Journal of Motor Behavior* 19, 486–517 (1987)

7. Malik, S., Ranjan, A., Balakrishnan, R.: Interacting with large displays from a distance with vision-tracked multi-finger gestural input. In: Proc. of UIST 2005, pp. 43–52 (2005)
8. Matejka, J., Grossman, T., Lo, J., Fitzmaurice, G.: The design and evaluation of multi-finger mouse emulation techniques. In: Proc. of CHI 2009, pp. 1073–1082 (2009)
9. Parker, J.K., Mandryk, R.L., Inkpen, K.M.: TractorBeam: seamless integration of local and remote pointing for tabletop displays. In: Proc. of GI 2005, pp. 33–40 (2005)
10. Tokoro, Y., Terada, T., Tsukamoto, M.: A pointing method using two accelerometers for wearable computing. In: Proc. of SAC 2009, 136–141 (2009)
11. Toney, A., Thomas, B.H.: Applying reach in direct manipulation user interfaces. In: Proc. of OzCHI 2006, pp. 393–396 (2006)
12. Yoshikawa, T., Shizuki, B., Tanaka, J.: HandyWidgets: local widgets pulled-out from hands. In: Proc. of ITS 2012, pp. 197–200 (2012)

Part III
Touch-Based Interaction

Human Centered Design Approach to Integrate Touch Screen in Future Aircraft Cockpits

Jérôme Barbé¹, Marion Wolff², and Régis Mollard²

¹ AIRBUS Operations SAS, 316 Route de Bayonne, BP 3153, 31060 Toulouse Cedex 09
jerome.barbe@airbus.com

² Univ. Paris Descartes & ESTIA/PEPSS Technopôle Izarbel, 64210 Bidart
{marion.wolff, regis.mollard}@parisdescartes.fr

Abstract. This research aimed at developing new types of Human-Machine interaction for future Airbus aircraft cockpit. Touch interaction needs to be studied because it brings some advantages for pilots. However, it is necessary to redefine pilot's workspace to optimize touch interaction according to pilot population characteristics and human physical capabilities. This paper presents the touch interaction area model and the tactile assessment carried out to validate our hypothesis, leading to rules/guidelines for cockpit layout and HMI designers.

Keywords: Human Centered Design, interaction design, anthropometry, touch screen interaction, guidelines.

1 Introduction

Many research studies are carried out at Airbus to address new types of Human-Machine interaction for pilots. This study was focused on the integration of touch screen technology in future aircraft cockpit. Indeed, this technology brings some advantages 1) for pilot interaction, for example [1]: intuitive operation requiring little thinking by a direct manipulation, easier hand-eye coordination, and 2) for cockpit definition: more software flexibility, space optimization with no extra workspace required for physical input devices such as command buttons or pointing devices. But touch screens need to be installed at a specific location inside the cockpit for accessibility and working postures concerns. Indeed, if not properly located, the use of touch screen may induce muscular fatigue, musculoskeletal disorders and could also degrade Human-Machine interaction [2], [3].

The aim of this study was to define digital manikins with their associated reach envelopes and postures to provide a design model to better integrate touch screens and to optimize Human-Machine interaction. The part of the study related to anthropomorphic and biomechanics needs, focusing on pilot characteristics and working postures to be considered for building up the design model were addressed in a former paper [4]. This paper will present the global design concept and the results of the tactile assessment taking into account criteria such as touch and task performance (gesture

accuracy, duration and difficulty of the task). This model provides rules and guidelines to be used by designers for optimizing touch screens technology integration in the cockpit. It also ensures an efficient Human-Machine interaction.

2 Theoretical Design Model

The theoretical design model is based on 6 modules (figure 1). The criteria related to Human-Machine Interaction and human physical capabilities were considered for formalizing the Touch Interaction Area Model. Each module is defined with its associated criteria.

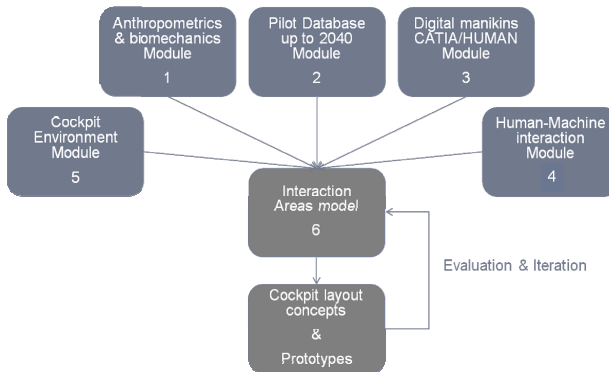


Fig. 1. Theoretical Design Model

2.1 The Anthropometrics and Biomechanics Module (1)

How to integrate human physical capabilities in the design of new equipments and systems to improve comfort, safety and efficiency? We focused mainly on the choice of the relevant anthropometric key measurements and the definition of working postures, reach and visual capabilities.

Sitting height and buttock knee length (related to seated positions) and forward reach were retained as key anthropometric measurements.

Working postures of pilots were defined according to angles of less discomfort, muscular efforts and energy expenditure reduction, contact pressures and internal disks pressures reduction. Four postures were chosen as reference. Their characteristics are: minimum discomfort and maintained during the performance of operational tasks and/or rest periods in the cockpit (“Theoretical Design Eye Posture (DEP)”, “Functional DEP”, “Monitoring – Cruise” and “Nap posture”). We also selected additional postures to be studied (reach areas extended) for short duration task. They are associated with bending and/or twisting of the trunk and movements of one or both upper limbs: « Forward seated posture », « Forward maximum », « Forward 45° » and « Upward » posture.

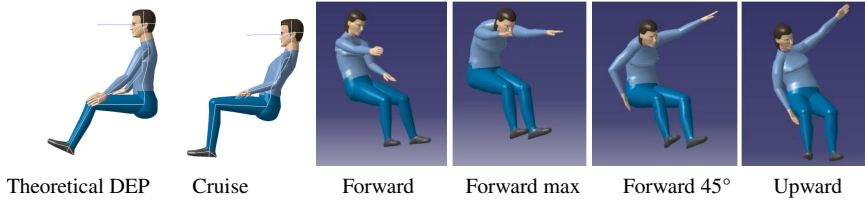


Fig. 2. Working and additional postures for manikins created on CATIA/HUMAN

Reach capabilities for upper limb is defined by the functional hand reach envelop consisting in fully gripping a rod with the hand (Forward Reach (FR) see ISO 7250). Allowances were added to increase the reach distances: for pinch grip, FR+50 mm (figure 3-1) and, for touch, FR+100 mm (figure 3-2). A preferential area was also defined for interaction with both hands (figure 3-3) [5].

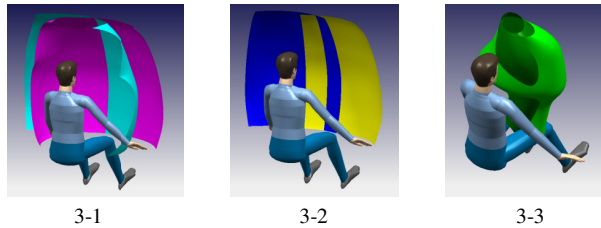


Fig. 3. Functional and preferential reach areas for left and right hands

Visual capabilities were also considered. The visual area modeling was based on the ISO 14738 to define the adequate field of vision (α) with or without movements of the head (β) and the body (γ) (figure 4).

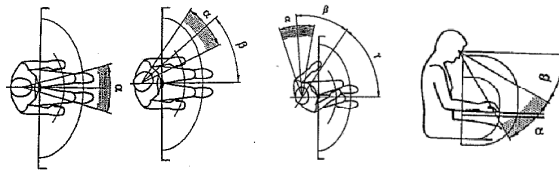


Fig. 4. Visual areas according to head and body movements (from ISO 14738)

2.2 The Pilot Database Module (2)

How to integrate morphological variability ensuring that both small and large pilots are able to reach and interact comfortably and efficiently with the touch screens from their seated position? We built up a pilot database (prediction up to 2040) taking into account the evolution of the morphologies and the geographical locations for key measurements. The range of the pilot population are defined from the small (5th percentile) Asian Japanese female to the large (95th percentile) European North male. Populations are defined by anthropometric surveys extracted from WEAR database systems [6].

2.3 The Digital Manikins Module (3)

How to perform ergonomic studies using Computer Aided Design (CAD) techniques? We chose manikins and defined body dimensions according to the variability of length for sitting height, buttock-knee length and forward reach. A set of « boundary » manikins were derived with different morphotypes related to trunk/lower limbs ratio, using bivariate distributions to define the appropriate range of variations for the 5% and the 95%. Sitting height was chosen as key dimension for 5%-95% variability and Forward reach was adjusted to cover the variability for the 2040 populations. Pair of manikins was created for each percentile retained: 5% and 95%.

2.4 The Human-Machine Interaction Module (4)

Which Human-Machine Interaction criteria to be considered?

- Duration of the tactile task (D): maintaining posture and hand motion. Duration of the tactile task may have a negative impact on comfort if duration is too long because the muscles have to fight against gravity effect to maintain the positions of the upper limbs.
- Frequency of the task (F): time for rest (sec) between two tactile tasks without upper limb rest on appropriate supports. The frequency of the task may induce fatigue at joint levels and musculo skeletal disorders (if too high).
- Repetitiveness of the task (R): repetition of similar task (in terms of gesture and posture constraints) during a flight. It contributes to physical fatigue at the postural and gestural levels.
- Gesture library (G): gestures selected for the touch screen interaction (Number of fingers & hands used).
- Task difficulty (T): accuracy and difficulty of the gesture to perform the task. They are the main factors of muscular and articular fatigue (necessity for maintaining a posture and oculomotor control of the gesture).

2.5 The Cockpit Environment Module (5)

How to consider environmental context and cockpit layout constraints? We developed the model to be used in several aircraft cockpit concepts. In this case, we took into account the specificities of the environmental context, such as the vibration (turbulences) and the cockpit layout constraints that could have an impact on touch interaction, for example: the seat position, the visual field, the cockpit nose dimension, the display locations (right /left) and the fact that information on central displays is shared by the two pilots. Moreover, lateralization (i.e. right or left hand used) needs to be addressed as it could affect the performance, the precision and the comfort of touch interaction.

2.6 Touch Interaction Area Model (6)

All the criteria from the different modules (table 1) have been gathered to build up the Touch Interaction Area Model in order to optimize the interaction according to the touch screen locations. For each criterion, we made hypotheses to be assessed.

Table 1. Criteria retained in the Touch Interaction Area Model

Type of criteria	Criteria
Human-Machine Interaction	Duration of the tactile task (D)
	Frequency of the task (F)
	Repetitiveness of the task (R)
	Gestures used (G)
	Task difficulty /accuracy (T)
Human physical capabilities	Working Posture (WP)
	Lateralization (L)
	Visual Control (VC)
	Vibration (V)

With these hypotheses, a cockpit concept was designed on CAD taking into account cockpit constraints (figure 5). Visual boundaries and reach adjustments (FR+50 mm and FR+100 mm) were based on the 5th percentile digital manikins.

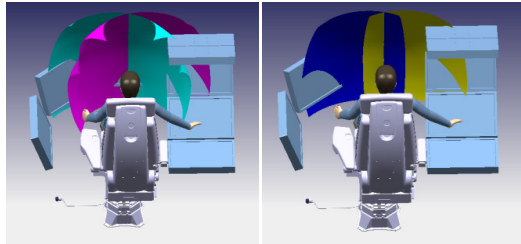


Fig. 5. Cockpit concept defined on CAD based on Touch Interaction Area Model

3 Method

Our experimental approach consisted in collecting data from a sample of 10 pilots in order to validate postures, reach area capabilities and touch interaction areas predefined with digital manikins in CAD. This should allow us to confirm the acceptability for both small and large pilots in terms of accessibility, postural constraints, physical efforts, task difficulty and visual boundaries.

3.1 Experimental Set up

We developed a physical mock-up (1/2 cockpit, at the right seated position) with 8 touch screens based on the CAD cockpit concept (figure 6, left). A Motion capture

tool (Moven system) was used to capture and analyze pilots' postures and functional reach envelopes in real time (figure 6, right). Anthropometric measurements for each subject were also integrated in Moven system to get personalized avatars.

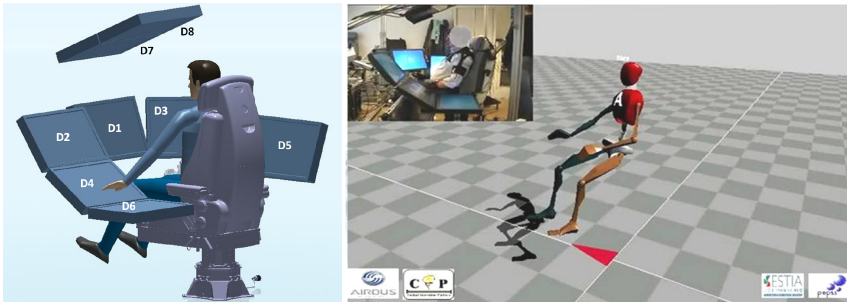


Fig. 6. Experimental mock-up and Moven capture

A sample of 10 right-hand pilots (8 males and 2 females) (age $m = 46.5$; $SD = 13.5$) was selected to cover morphological diversity for key measurements (Sitting height and Forward reach) and the pilot population, as defined in the pilot database module (§2.2).

3.2 Experimental Tasks and Protocol

Two types of tactile tasks were addressed during assessment:

- Accurate tasks: typing a short text on a virtual keyboard, selection of a waypoint on a Flight Management System (FMS) prototype or creation of a puzzle,
- Non accurate tasks: manipulation of a map to select an airport view or charts.

We collected postures and upper limbs movements when subjects performed tactile tasks on different display locations (figure 6). Six positions were analyzed (D1 to D6) with three tasks performed per display. All tests were carried out at the right seated position. A subjective assessment for the following 4 dimensions was used at the end of each task to collect the perceived level of: the performance (accuracy + quickness), the difficulty of the tactile gesture (hand motion), the physical effort (postural constraint) and the acceptability of the task (gesture + posture). Visual issues and right/left hand preferences were also collected.

Two kinds of functional postures were already characterized in a former paper [4]: postures with or without (or only small) constraints. With no constraints, the head and the pelvis are in neutral position. The lumbar has no postural constraint but there is no rest for the thoracic cage and forward upper limbs. Elbow and shoulder angles are in the range of less discomfort. On the contrary, with constraints, the head is at the acceptable limits for rotation and flexion angles; the trunk is highly constrained. So are the shoulder and the wrist extension angles.

Subjective assessments on physical effort (display location, right/left hand preferences and visual aspects) were also addressed in this former paper [4].

For this paper, we focused on the following Human-Machine interaction hypotheses to identify the relations between postures, task duration, display location and acceptability of the tactile tasks:

- H1: Postural constraints depend on display location,
- H2: Postural constraints and task duration have an influence on subjective assessments,
- H3: Subjective assessments vary according to attributed tasks, postural constraints and display location.

In order to verify these hypotheses a statistical analysis (descriptive and inferential statistics followed by geometric data analysis) [7], [8], [9] was conducted on the following data: subjective assessment, tasks (with and without accuracy), displays (D1 frontal, D2-D4 central, D3 lateral, D5-D6 backward), postures (with and without constraints), and task duration: short (9-15 s), medium (16-60 s), long (> 60 s). The resulting table was constituted of 178 lines (10 pilots x 18 tests minus 2 missing tests) and of 5 columns (4 subjective assessments and task duration).

4 Results

• Postural constraints depend of display location. (H1)

Cross data analysis (Posture/Displays) for subjective assessments on performance shows that pilots make the choice of postural constraints to be more efficient when displays are located in a central position ($t [58] = 2.24$; $p < .003$). Right-handed subjects prefer both hands with postural constraints (physical effort significantly higher) than their left hand to perform tactile tasks on central displays (effect of lateralization). The choice of such postural constraints contributes also to increase legibility and to lower parallax difficulties (presented in former paper [4]). It is also explained by the need for the subjects to get a better field of vision as recommended by ISO norm 14738 (see 2.1).

For other subjective assessments, tactile task difficulty is low for frontal/lateral displays ($t [92] = 7.27$; $p < .0000$) and increases for the other displays (central and backward); the difficulty is the higher for central displays when pilots adopt a posture without constraint ($m = 43.94$; $SD = 21.06$). Acceptability is better for frontal/lateral displays for “no constraint condition” than for the others. This acceptability for the others remains at a level greater than 50% (results differ significantly with central display: $t [92] = 4.42$; $p < .0000$).

• Postural constraints and task duration have an influence on subjective assessments. (H2)

The performance is better perceived for short and medium tasks duration than long tasks whatever postural constraints (with constraints: $t [82] = 2.52$; $p < .01$, without constraints: $t [92] = 4.84$; $p < .0000$). In addition, task difficulty increases with long duration (with constraints: $t [82] = 2.10$; $p < .04$ - no constraint: $t [92] = 3.76$; $.0003$). Acceptability remains correct whatever task duration (no significant results).

- **Subjective assessments vary according to attributed tasks, postural constraints and display location. (H3)**

The correlations between the 4 subjective assessments dimensions and duration were carried out on the sample of the collected data (table 2). These correlations were projected in a geometric space (vectors/variables space). All correlations are significant. Performance has a negative correlation with task difficulty, physical effort and duration and a positive correlation with acceptability.

Table 2. Correlation matrix

Variables	Performance	Task Difficulty	Phys. effort	Acceptability	Duration
Performance	1.00	-0.65	-0.46	0.61	-0.43
Task Difficulty	-0.65	1.00	0.72	-0.59	0.40
Phys. Effort	-0.46	0.72	1.00	-0.62	0.22
Acceptability	0.61	-0.59	-0.62	1.00	-0.18
Duration	-0.43	0.40	0.22	-0.18	1.00

For the Principal Component Analysis (PCA), duration was not retained because previously changed to a qualitative variable. All the quantitative variables were indexed with 8 qualitative variables (task identification, type of display, characterization of the posture, visual constraint, hand lateralization, task categorization, forward reach and duration). These qualitative variables allowed us characterizing the behavior profiles in the “clouds” derived from PCA (derived clouds of mean points).

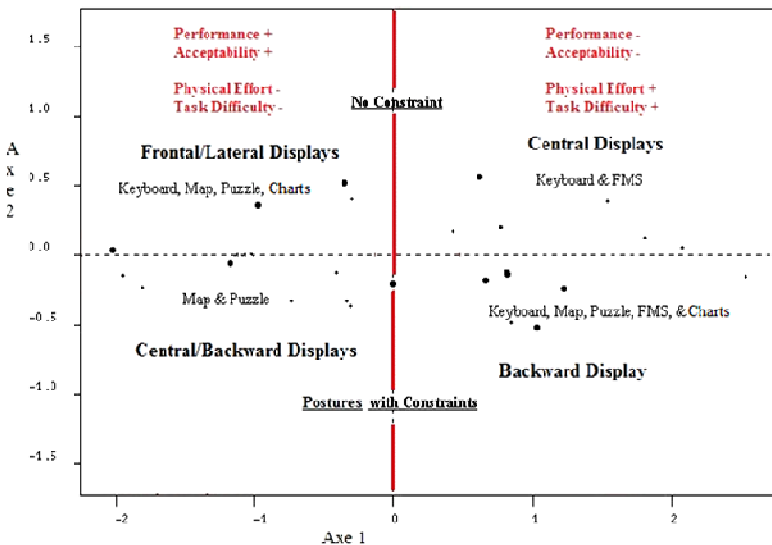


Fig. 7. Derived cloud of the posture/display/task mean points

PCA consists in building up a vector space of variables and a Euclidean space of individuals' points. Graphical representation (figure 7) summarizes results obtained from interpretation of vectors-variables' projection on axes 1 and 2 (84% of variance accounted) and from different configurations of individuals' mean points indexed with qualitative parameters (postures with or without constraints / display locations / type of tasks). Axis 1 describes opposition between performance/acceptability (figure 7, left +; right -) and physical effort/task difficulty (figure 7, left-; right +). Axis 2 represents the opposition between unconstrained (top) and constrained (down) postures. Best scores for performance/acceptability (left) are for frontal and lateral displays (keyboard, map, puzzle, Charts) and also for central and backward displays (map and puzzle). Opposite (right), backward (keyboard, map, puzzle, FMS, Charts) and central displays (keyboard and FMS) are less acceptable.

5 Preliminary Guidelines

Based on the results presented in this paper and in the former one [4], four categories of interactions have been defined according to the criteria of the Touch Interaction Area Model (table3). Some preliminary guidelines can be extracted from this applied model, for example: tasks with a high level of difficulty and gesture accuracy are to be performed preferentially on frontal or lateral displays; simultaneous action of both hands is possible but not recommended for some display locations due to postural constraints; if display locations imply upper limb elevation and postural constraints, duration should not be long due to muscular fatigue.

Table 3. Categories of interaction according to the touch screen locations

Type of criteria	Criteria	CAT1	CAT2	CAT3	CAT4
Human-Machine Interaction	Duration of the tactile task (D)	Long	Medium	Short	One time
	Frequency of the task (F)	Not tested			
	Repetitiveness of the task (R)	Not tested			
	Gestures and hands used (G)	All gestures	Gestures with both hand possible	One hand limited	Mono-touch only
	Task difficulty /accuracy (T)	Accuracy	Accuracy	No accuracy	No accuracy
Human physical capabilities	Working Posture (WP)	Preferential area	FR+50 with small constraints	FR+50 with constraints	FR+100 All postures
	Lateralization impact (L)	Low	High	Medium	Low
	Visual Control (VC) (field of vision & head movement)	α (30°) $\alpha+\beta$ (60°)	$\alpha/2+\beta$ (55°) $\alpha+\beta+\gamma$ (90°)	δ (90°) $2\alpha+\beta+\gamma$ (110°)	All degree of freedom
	Vibration impact (V)	Not tested			
	Touch screens location	Location not tested	Frontal/Lateral (D1, D3)	Central (D2, D4)	Backward (D5, D6)

This preliminary definition of the categories answers two types of question that can be addressed by cockpit layout and HMI designers:

- Where should the HMI be localized according to the type of interaction needs? Answer: If the HMI interaction requires: G = one hand gesture, T = gesture

accuracy and D = medium duration, then the touch interaction area should be CAT2 (frontal or lateral) location.

- What types of interaction should be recommended for a system in a dedicated touch screen location?

Answer: If the HMI is located on the central area, then interaction should be CAT3 (D = short duration, T = no gesture accuracy and G = limited to one hand gesture).

6 Conclusion

The Touch Interaction Area Model has proven to be worthy to define and give guidelines for design. Nevertheless, complementary studies with real operational tasks are needed to validate these new ergonomic rules in order to refine the nature of the tactile task, the gesture accuracy and also to study the impact of the frequency and the repetitiveness of the task in the model. Vibration effects such as those encountered in turbulence conditions need to be investigated to identify postural, upper limb and Human-Machine interaction disturbances. It will also be interesting to study if, an adapted training improves the way pilots interact with touch screen technology.

References

1. Shneiderman, B.: Touchscreens now offer compelling uses. In: Sparks of Innovation in Human-Computer Interaction. Ablex Publ., Norwood (1993)
2. Young, J.-G., Trudeau, M., Odell, D., Marinelli, K., Dennerlein, J.-T.: Touch-screen tablet user configurations and case-supported tilt affect head and neck flexion angles. *Work* 41, 81–91 (2012)
3. Fuller, H., Tsimhoni, O., Reed, M.P.: Effect of In-Vehicle Touch Screen Position on Driver Performance. Proceedings of the Human Factors and Ergonomics Society Annual Meeting 52, 1893 (2008)
4. Barbé, J., Chatrenet, N., Mollard, R., Bérard, P., Wolff, M.: Physical ergonomics approach for touch screen interaction in an aircraft cockpit. In: Proceedings of the Ergo'IHM 2012 Conference. ACM, New York (2012)
5. Ignazi, G., Mollard, R., Pineau, J.C., Coblenz, A.: Reconstitution en trois dimensions des aires d'atteintes du membre supérieur à partir de quelques données biométriques classiques. *Cahiers d'Anthropologie* 3, 93–117 (1979)
6. Mollard, R., Ressler, S., Robinette, K.: Database contents, structure, and ontology for World Engineering Anthropometry Resource - WEAR. In: Proceedings of the 16th Triennial World Conference of the International Ergonomics Association, July 10-14. The Netherlands, Maastricht (2006)
7. Benzécri, J.P.: Correspondence analysis handbook (Benzécri, J.P. Trans.). New-York: Dekker (Original Work published 1980) (1992)
8. Wolff, M.: Apports de l'analyse géométrique des données pour l'analyse de l'activité. In: Sperandio, J.-C., Wolff, M. (eds.) Formalismes de Modélisation Pour l'analyse du Travail et l'ergonomie, pp. 195–227. PUF, Paris (2003)
9. Le Roux, B., Rouanet, H.: Geometric Data Analysis. Kluwer Academic Publishers, Dordrecht (2004)

Evaluating Devices and Navigation Tools in 3D Environments

Marcela Câmara, Priscilla Fonseca de Abreu Braz, Ingrid Monteiro,
Alberto Raposo, and Simone Diniz Junqueira Barbosa

Departamento de Informática, PUC-Rio
Rua Marquês de São Vicente 225 – 22451-900 Rio de Janeiro, RJ - Brazil
{mcamara, pbraz, imonteiro, abraposo, simone}@inf.puc-rio.br

Abstract. 3D environments have been used in many applications. Besides the use of keyboard and mouse, best suited for desktop environments, other devices emerged for specific use in immersive environments. The lack of standardization in the use and in the control mapping of these devices makes the design task more challenging. We performed an exploratory study involving beginners and advanced users in the use of three devices in 3D environments: Keyboard-Mouse, Wiimote and Flystick. The navigation in this kind of environment is done through three tools: Fly, Examine and Walk. The study results showed how the interaction in virtual reality environments is affected by the navigation mechanism, the device, and the user's previous experience. The results may be used to inform the future design of virtual reality environments.

Keywords: 3D environments, evaluation, navigation tools, user experience.

1 Introduction

3D environments have been increasingly used in several applications. In order to make the navigation easier, many devices specific to immersive environments emerged, to substitute or complement the keyboard and mouse, which are best suited for desktop environments. The lack of conventions in the control mapping of these devices makes the design of 3D environments quite challenging.

In addition to the challenge related to the physical use of these devices, we have to face challenges related to the interaction based on 3D interfaces, which are more complex than those based on WIMP bi-dimensional interfaces. This happens because the latter offers a synthetic view of interaction possibilities, progressively brought to the user through a clearly planned and understandable sequence of windows and panels. Because of a host of established conventions, the user generally knows which sequences of actions perform some wanted operation. Equivalent results may be obtained in different ways involving different interaction styles, but the number of alternative behaviors is usually small, *e.g.*, menu selection vs. keyboard shortcuts. Conversely, the interaction in 3D environments involves an exploratory approach and requires typical real world operations like moving, navigating around objects, and so on. For each step, there are several actions and many ways of alternating movements during the interaction with these devices [2].

Another issue that goes beyond the interface and the devices mapping is related to the user's understanding while navigating in a 3D environment. In order to navigate in a satisfactory way, some aspects have to be considered, such as the sense of orientation. Zhai et al. [9] talk about the orientation sense as an important point, describing users who are "out of their goals" or who "lose their location". Furthermore, Robertson et al. [5] point out the difficulty of "having the sense about where you are or knowing what is behind you". This can clearly affect the ability of finding information and performing the task efficiently.

With the goal of investigating the challenges related to the navigation in 3D environments, we did a study with beginner and advanced users during the use of three devices: Keyboard-Mouse, Wiimote and Flystick, relating their use with three widely used navigation mechanisms: Fly (on the scene), Examine (the scene around a specified point) and Walk (in a surface). The way the user deals with these mechanisms depends on the chosen interaction device. The results of this study showed how the use of the devices, together with the users' experience and the knowledge and use of the three navigation mechanisms affect the interaction in virtual reality environments. Furthermore, we identified problems and obtained suggestions from the participants that may be useful for the design of 3D applications that make use of these devices and tools.

2 The LVRL Framework

LVRL (*Lightweight Virtual Reality Libraries*) is a non-intrusive framework that allows the conversion of desktop applications in immersive ones, in a way that both types of environments (desktop and immersive) : in a way that both type of environments can be interchanged at execution time. Regarding the output, the difference between an immersive application and a desktop application is related to the fact that the first application supports multiple video outputs from distinct viewpoints [7]. Regarding the input, the main difference between these applications is that, in immersive environments, one should use "non-conventional" interaction devices and techniques.

Three navigation mechanisms were used to evaluate the framework in a 3D environment: Fly, Examine and Walk. *Fly* allows the camera to fly through the scene at a given speed and the user to freely exploit the environment. *Examine* allows an object or location in the scene to be inspected. Its operation consists of rotating the camera around a point of interest, called rotation center or pivot, also allowing to zoom in on this object. To do so, it is necessary to first choose the object that will be the pivot (number 4 in Fig. 1). If the pivot is poorly specified, the resulting behavior is likely to confuse the user [3]. *Walk* allows the user to walk around the scene. It is similar to Fly, but now gravity applies, giving the feeling of really walking on a surface. During navigation, it is possible to set the speed of Fly and Walk.

The investigated environment (Fig. 1) has a toolbar with icons corresponding to the three kinds of navigation (identified by the number 2 in the image): Fly, Examine and Walk, to the pivot setting (number 3) and other options not addressed in our study. The option to go back to the initial position of the scene is located in the pull-down

menu (File → Reset). Using only Keyboard-Mouse, users can navigate and directly manipulate objects at the user interface and the camera. With the other devices, these operations are done through dedicated buttons in the device.

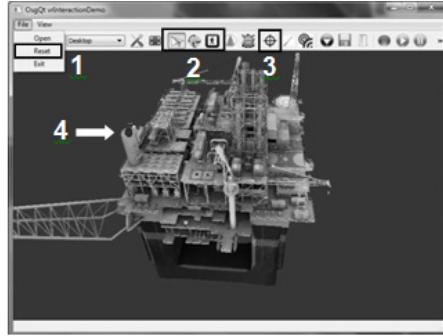


Fig. 1. The environment user interface

3 Evaluated Devices

We evaluated three kinds of devices in our study. The first one is **Keyboard-Mouse**, whose mapping to the navigation mechanisms is described in Table 1.

Table 1. Keyboard-mouse mapping

		Fly	Walk	Examine
Key-board	Arrows	To move to front, back, left, right		-
	A	To move to up	-	-
	Z	To move to down	-	-
Mouse	Scroll	Speed control		Zoom
	Click	Keeping the button pressed and dragging the mouse, you orient the movement direction.		To select pivot
	Drag			Camera rotation around the pivot

The second device is **Wiimote**, a 3D input device of Nintendo Wii game console. It brings a series of buttons to communicate with the console and, to detect movement, it has an accelerometer and an infrared sensor. Wiimote became a device used by 3D application developers because it requires a single Bluetooth receptor.

In the LVRL framework, Wiimote movements are captured only with the accelerometer. In addition to the device buttons, only two movements are supported: “pitch”, which is the rotation movement upon the transversal axis; and “roll”, which is the rotation movement upon the longitudinal axis. However, there is a movement, called “yaw”, which is not adopted with Wiimote in the LVRL framework, due to the lack of support for the infrared sensor, in charge of recognizing this movement. Table 2 summarizes the Wiimote–framework mappings, and Fig. 2a shows the Wiimote and the layout of its buttons.

Table 2. Wiimote mapping

		Fly	Walk	Examine
Buttons	I	To shift the navigation mode		
	Home	Camera initial position		
	Arrows	To move to front, back, left, right	Camera rotation around the pivot	
	A	-		To set and select pivot
	+ and -	Speed control		Zoom
Pitch and Roll		To orient the movement direction		-

Table 3. Flystick mapping

		Fly	Examine	Walk
Buttons	Analogic	Move	To move to front, back, left, right	Camera rotation around the pivot
		Click	-	Set and select pivot
	B1	Camera initial position		
	B2 and B3	Speed control	Zoom	As in Fly
	B4	To shift the navigation mode		
	Pitch, Roll and Yaw		To orient of the movement direction	-

The third device evaluated is **Flystick**, a wireless interaction device developed by ART Tracking [1] for virtual reality applications. It has six buttons and an analogic directional button. The Flystick movement recognition is done through two infrared cameras placed in opposite sides, one in the right and the other one in the left of the projection. The orientation of movement direction, in Fly and Walk modes, depends on the movement of the hand when the device trigger is pressed. Flystick supports the pitch, roll and yaw movements. Table 3 presents the mapping of Flystick buttons. As seen in Fig. 3, Flystick has four non-labeled buttons aligned below and around the directional one. In the table, from left to right, they are identified as B(1-4).



Fig. 2. (a) Wiimote device. (b) Flystick device.

4 User Study: Participants and Procedures

We conducted an exploratory study with users, aiming to investigate the users’ perception of the LVRL-device mappings and to collect the users’ opinions about the

three devices, following a qualitative research approach [4]. In our study, potential users of an application developed with the LVRL framework navigated through a 3D model of an oil platform, where they should execute the proposed task. Our main objective was to investigate the use of three devices, considering the mapping of the different controls determined by the framework developers. We observed how people interact with the involved devices and how they understand the control mappings. In addition, we tried to identify their difficulties, preferences and suggestions for improvement of the devices' use and navigation mechanisms.

We determined the users' profile based on the framework features involved in this research. Six right-handed participants were recruited, categorized in two profiles: beginner and advanced. The participants are identified by PB1, PB2 and PB3 for beginners and PA1, PA2 and PA3 for advanced ones. The beginners were 26, 35 and 36-year old women who did not have previous experience with devices in 3D environments. The advanced users were 25, 27 and 31-year old men who use 3D applications at least once a day. All of them had previous experience with 3D visualization and 3D games. All of them had used game devices such as Wiimote, Joystick and Kinect at least once. PA3 was the only who had used the Flystick beforehand.

The test was divided in five stages: 1) introduction and application of pre-test questionnaire; 2) explanation about the framework and presentation about the device mapping; 3) training with the device; 4) task execution; 5) semi-structured interview.

The users executed the same task using Keyboard-Mouse, Wiimote and Flystick. In the proposed task, they should navigate on an oil platform (as seen on Fig. 1) using the navigation tools: Fly, Walk and Examine, according to the presented instructions. In each step of the task we suggested them to use a specific tool, but they could use the preferred one. They executed the task using each device separately following a pre-determined order. The order of the used devices was modified between the users to reduce learning effects in the study results. After the execution of each task, a brief semi-structured interview was conducted to capture the user's opinion about the mapping of the used device.

5 User Study: Results Discussion

5.1 Navigation Tools

We present data related to the users' interpretation, use and preference regarding the three navigation tools. The evidences reported here can be generalized to the tool concepts themselves. Even when the issue occurred during the use of a specific device, we have noticed that it is also applicable to the other devices.

Fly Mode. This navigation mode was, in general, understood by the participants. It was also one of the most used, due to its flexible navigation. Among the beginners, we noticed a frequent use of Fly to recover from falls during the Walk mode. Some problems occurred during Fly. For example, PB1 tried to use zoom, available only in Examine. PB2 tried to "spin" around a point in the platform with Fly, instead of the more recommended tool, Examine. This participant thought that Examine was just to

select the pivot and then, to turn around the vision, it would be necessary to use Fly. Regarding the advanced participants, PA1 complained about this mode. He would like to see options such as zoom in and zoom out for Fly, as it is possible with Examine.

Examine Mode. Although Examine has a very specific function and fewer handling options in comparison with the other two modes, both user groups had some difficulties in using it. All problems identified were related to marking the pivot. PB2 and PA1 had trouble recognizing that the object was already marked as pivot. In other words, they believed that, whenever they used Examine, it would be necessary to mark the pivot, when in fact, you can use Examine using a previously defined pivot.

Walk Mode. This navigation tool presented the greatest difficulty to both groups. The main reason for this is the peculiar feature of the Walk Mode: locomotion on surfaces only. When the user approaches an “open” area, he suffers the effects of gravity and starts to fall until he or she finds another surface. This feature, though realistic, caused ample frustration among participants. Because of this problem, during task performance, although recommended to use Walk, participants preferred to repeatedly use Fly. We could then identify three common situations that caused falls during Walk: (1) When changing from Examine to Fly, the users must pass by Walk. If they were not on a surface, the user fell. The way to overcome this problem was to make trading so quickly so as not to give time for the selection of Walk to take effect. (2) There is a lack of peripheral vision in Walk, so it is hard to see the boundary surfaces in some situations, especially at the sides, since usually the camera placement is forward. (3) There was a difficulty in understanding what a safe surface is. While in Fly or Examine mode, when participants changed to Walk, they sometimes did not realize that there was no surface directly below them.

PB3 gave us an interesting suggestion to work around this falling problem: enable Walk mode only when you are in a favorable position above a safe surface.

5.2 Interaction Devices

In this section we present and discuss data that specifically address each device, considering issues related to handling and mapping.

Keyboard-Mouse. Most participants reported problems related to the sensitivity of the devices. In all cases, the movements carried out using the mouse or keys resulted in very fast movements on the display, thereby exposing the high sensitivity of this device. A consequence of this problem was that PB1 and PB2 strongly avoided using the mouse. They were trying to get where they wanted by using only the arrow keys. They used the mouse only in Examine (which has no associated function on the keyboard). However, unlike them, PB3, even having reported the problem of sensitivity, used the mouse quite often. PA2 also complained about the mouse high sensitivity.

PB1 and PB2 reported that they liked to use Examine with Keyboard-Mouse. We attribute this preference mainly because the pivot was marked with just the click of the mouse, unlike other devices, where it is selected with a virtual ray pointed in the screen. Both users stated that the use of the mouse was easier only with Examine. PA1 was pleased with the mouse sensitivity, and did not need to control the speed any

time, neither in Walk nor Fly. The advanced participants, in general, did not show as many problems as the beginners. PA1 and PA2, for example, performed the task very fast, with only a very brief navigation. PA1 also dismissed the initial training time, starting directly to the task. He, after all, considered that he had an “obvious” facility using mouse and keyboard in 3D environments.

Regarding the mapping of the controls on this device, we have seen problems related mainly to the use of the mouse wheel (used to control the speed). PB1, PB2 and PA2 did not like this mapping. PB1 and PB2 agreed that this speed control was the worst problem of using Fly with Keyboard-Mouse. Another problem observed was that PB2 thought she could rotate the camera using the directional keypad. After a while, she realized that it was the mouse that controlled the camera. These participants felt bad having to use the mouse and keyboard at the same time. They would like to do everything on the keyboard. PB2 explained that for navigation tools Fly and Walk, she preferred directional commands that could be mapped to the far left of the keyboard, rather than the arrows. Thus, it would be best to use your left hand to set the direction of movement. She also suggested that the directional arrows were used to perform the movement of the camera, originally mapped to the mouse.

PA1 and PA2, who had experience with keyboard and mouse in 3D environments and / or games, suggested that the functions of the directional arrows should be mapped onto the keys “W”, “A”, “S”, “D”, which is a common pattern in computer games. PA2 also suggested that the functions keys could be kept to those who preferred to use that way. Another option was also to replicate the functions of the mouse on the arrows, for those who wanted to do everything from the keyboard.

Wiimote. While PB1, PA1 and PA2 used only the right hand to handle the device; PB2, PB3 and PA3 preferred to use both hands to change the navigation mode and to increase or to reduce speed, as a way to streamline and facilitate their interaction with the device.

Among all participants, the most recurrent problem was the limitation imposed by the device relative to movements to the left and to the right. In this case, participants should rotate the control to the sides (roll, previously discussed), while the upward and downward motion was to raise the control to these directions (pitch). In many situations, PB1 and PB2 moved the hand sideways (yaw) instead of rotating it (roll). PA2, PB3 and PA3 said it was more comfortable and natural moving side to side (like in Flystick) instead of rotating the control. Moreover, PA2 explained that the turning motion had little precision, which threatened his locomotion.

Another common problem was the lack of a rest position during Walk. Even if the participant did not move the control, the camera did small and constant movements and caused unwanted displacements. Because of this, PB3 used both hands few times to help the movement. To solve this problem, PA1 suggested using the trigger to “lock” movement. PA1 reported that he did not consider practical the speed control of Wiimote and PB1 reported that she avoided using this function. She would prefer going slowly, feeling control of the situation because she was afraid of getting lost.

The participant PB1 had difficulty to select the pivot and she reported there was no precision in the Wiimote movement. Except for selecting the pivot, she was satisfied

with using Examine. PB2 suggested that the exchange of navigation tools could be performed by the button “B” (trigger) rather than with the button “1”. According to the participant, this suggestion was based on the location and accessibility of buttons (Fig. 2). Some participants reported a problem related to the shift of these tools by a single button on a cyclical basis, mainly due to the obligatory passage by Walk when this was not the goal. PB1 and PB3 suggested that were used different buttons for navigating to the right and to the left among the navigation tools. PA1 and PA2 suggested that Walk could be activated by a separated button.

Flystick. The beginners demonstrated greater acceptance of the Flystick device. In some occasions, PB1 highlighted its ease of use; she liked the ergonomic characteristics of the device and the way to handle it. She said: “My perception is that I can map better my intention with the movement of this device”. PB2 also liked the good control response. PB1 and PB3 also praised the analog control device.

Some participants reported problems related to the sensitivity of the hand movement (camera control) and of the analogic motion control (steering control). About the first case, PB3 and PA2 found that the sensitivity of the device was low, *i.e.*, the navigation was slow and imprecise. PB3 thought the camera rotated very slowly when compared to the physical movement that the device performed. PA2 used to raise the control abruptly and maintained it pointing upwards, while on the scene the camera lifted slowly. On the analogic control, PA1 and PA3 considered that their sensitivity was high during navigation in the Examine and Walk mode, respectively.

In addition, PA2 explained that using Flystick it was harder to turn corners during Walk and he would prefer to go through the middle of the platform to avoid falling.

When we asked PA3 what he thought most difficult to do with Flystick, he answered: “The hardest part was hitting the Walk mode, because you have the freedom to point to where you want to go is great, except that when you apply speed it becomes too fast. Sometimes you lose a bit of control”. Still regarding Walk, PA1 complained about not being able to lift his head up to look up. One of the few complaints about the Flystick mapping among the beginners was related to the combined use of hand movement to control the camera with the trigger activation: the camera only moved when the trigger was pressed. In some cases, this requirement meant that the participants moved the hand vigorously without viewing any result on the screen. In order to change the cyclic shifting between navigation modes, PB3 and PA1 suggested holding the button on the far right to navigate to the right and include the leftmost button to navigate to the left. PA2 suggested using a separate button for Walk, just as he had suggested for Wiimote. Regarding mapping suggestions, PA2 was emphatic: “So, my suggestion is to give it up”. He really did not like the handling of the device, but he had no complaints about the mapping.

Comparing the Three Devices. Regarding Keyboard-Mouse, PA1 considered this device “lighter” than Flystick. PB3 said she preferred Keyboard-Mouse because “he was used to it”. She said: “Well, I have more ease in using keyboard and mouse. I’m not used to playing or doing anything with that (*pointing to Flystick*)”. For PA3, if he did not have the option of keyboard and mouse, he would choose Flystick.

Considering Wiimote, PB2 and PB3 thought the speed control in this device was better than in the mouse. PA3 considered the speed control better in Wiimote than in Flystick. For PA1, Wiimote's motion and response is better than Flystick's. In contrast, he thought the accuracy in Wiimote worse than in Flystick due to the absence of yaw movement. He added that Wiimote is ergonomically worse than in Flystick. PB1 also cited this relation, by adding that the former had no advantage over the latter.

With respect to Flystick, among the beginners, this device has more advantages than the others. PB3, for example, said that the mapping of Flystick buttons was better than in Wiimote. The participant expressed her admiration for Flystick: "I really liked this analogic control. I found it much easier. It arrives fast where you want." PB2 stated that the camera movement in Flystick is better than in Keyboard-Mouse. PA3 said the speed control on Wiimote was better, but moving with Flystick was more natural. PA2 demonstrated a negative opinion regarding Flystick. When starting the test, he said it was very odd the use of Flystick and very different from Wiimote. Comparing Flystick with Wiimote, he recognized that the first one does not have the problem of moving the camera during Walk. But this problem was smaller than the bad movement of Flystick. He said: "The learning curve (of Flystick) is much higher than the other two".

In order to improve the highlighted viewing of preferences in this section, jointly with previously pointed evidences, Table 4 presents a global ranking of preferred devices for each participant, and indicates the use order of each device to facilitate comparison (the preferred device received score 1). The table reveals an overall rejection of Wiimote, which was not ranked last only by PA2. Keyboard-Mouse proved to be the preferred device among advanced participants. We believe this was due to the fact that they are used to it in 3D applications.

Flystick was in third place only in one case, and it was the preferred among beginners. We attribute this behavior to the Flystick particular feature of responding in the screen to real movements of the hand. The camera should be positioned, in the visualization, to where Flystick is pointed. This mapping between the physical and virtual behavior is much more direct than the other two devices.

Table 4. Use and preference order

	Keyboard-Mouse		Wiimote		Flystick	
	U	P	U	P	U	P
PB1	3	2	1	3	2	1
PB2	2	2	3	3	1	1
PB3	1	1	2	3	3	2
PA1	2	1	3	3	1	2
PA2	1	1	2	2	3	3
PA3	3	1	1	3	2	2

U = Use Order, P = Preference Order

6 Final Considerations

The lack of standardization in the use and control mapping of interaction devices in virtual environments makes the design of these environments a challenging task. Due to the importance of exploring aspects of the use of such devices, the current study

consisted in exploring the use of three devices in 3D environments: Keyboard-Mouse, Wiimote and Flystick with beginner and advanced users. The type of interaction in the 3D environment depends greatly on the chosen interaction device. From the analysis of the results, we observed that advanced users have very divergent opinion of beginners, which is quite common in 3DUI [8] [6]. In the studied case, the divergence of opinions proved to be stronger when comparing the devices, especially regarding the Flystick, which was the preferred of most beginners and criticized by the advanced participants, who preferred the Keyboard-Mouse. We know that users of 3D games are used to Keyboard-Mouse, which may have made its use seem more natural. With regard to beginners, without previous experience, the more natural may have been one of the devices that were designed specifically for an immersive environment. Should the user experience be decisive in 3DUI? If so, should it be more crucial than in WIMP? If the opinion tends to change with experience, how to design 3DUI? These questions show challenges that still need to be studied further in HCI.

Another finding was about Wiimote, which was the most criticized of all devices. Problems such as ergonomics and imprecise movements were the most cited. The only participant (advanced) who praised this device reported having good experience using it in games. But although he liked this device, he said that some changes should be made to improve it. We also hope to make a small contribution to the design of virtual environments, highlighting their interaction design challenges.

References

1. ART Tracking, <http://www.ar-tracking.com/> (accessed on: June 21, 2012)
2. Augusto, C., Pittarello, F.: Observing and adapting user behavior in navigational 3D interfaces. In: AVI 2004: Proceedings of the Working Conference on Advanced Visual Interfaces, pp. 275–282. ACM Press, New York (2004)
3. Fitzmaurice, G., Matejka, J., Mordatch, I., Khan, A., Kurtenbach, G.: Safe 3D navigation. In: I3D 2008: Proceedings of the 2008 Symposium on Interactive 3D Graphics and Games, pp. 7–15. ACM, New York (2008)
4. Lazar, J., Feng, J.H., Hochheiser, H.: Research methods in human-computer interaction. Wiley, New York (2010)
5. Robertson, G., Czerwinski, M., Dantzich van, M.: Immersion in desktop virtual reality. In: Proceedings of the ACM Symposium on User Interface Software and Technology, Banff, Alberta, Canada (1997)
6. Teather, R.J.: Comparing 2D and 3D direct manipulation interfaces. 112 p. Thesis of Master of Science, York University, Toronto (2008)
7. Teixeira, L., Trindade, D., Loaiza, M., Carvalho, F., Raposo, A., Santos, I.: A VR Framework for Desktop Applications. In: XIV Symposium on Virtual and Augmented Reality. XIV Symposium on Virtual and Augmented Reality – SVR 2012 (CD-ROM), Niteroi, RJ, Brasil (2012)
8. Trindade, D.R., Raposo, A.B.: Improving 3D navigation in multiscale environments using cubemap-based techniques. In: SAC 2011 – Proceedings of the 2011 ACM Symposium on Applied Computing, Taichung, Taiwan, pp. 1215–1221 (2011)
9. Zhai, S., Kandogan, E., Smith, B., Sekler, T.: Search of the ‘Magic Carpet’: Design and Experimentation of a Bimanual 3D Navigation Interface. *Journal of Visual Languages and Computing* 10, 3–17 (1999)

Computational Cognitive Modeling of Touch and Gesture on Mobile Multitouch Devices: Applications and Challenges for Existing Theory

Kristen K. Greene^{1,*}, Franklin P. Tamborello², and Ross J. Micheals^{1,*}

¹ National Institute of Standards and Technology, Gaithersburg, MD
{kristen.greene, ross.micheals}@nist.gov

² Cogscent LLC Washington, DC
Frank.tamborello@cogscent.com

Abstract. As technology continues to evolve, so too must our modeling and simulation techniques. While formal engineering models of cognitive and perceptual-motor processes are well-developed and extensively validated in the traditional desktop computing environment, their application in the new mobile computing environment is far less mature. ACT-Touch, an extension of the ACT-R 6 (Adaptive Control of Thought-Rational) cognitive architecture, seeks to enable new methods for modeling touch and gesture in today's mobile computing environment. The current objective, the addition of new ACT-R interaction command vocabulary, is a critical first-step to support modeling users' multitouch gestural inputs with greater fidelity and precision. Immediate practical application and validation challenges are discussed, along with a proposed path forward for the larger modeling community to better measure, understand, and predict human performance in today's increasingly complex interaction landscape.

Keywords: ACT-R, ACT-Touch, cognitive architectures, touch and gesture, computational cognitive modeling, modeling and simulation, movement vocabulary, gestural input, mobile handheld devices, multitouch tablets, model validation, Fitts' Law.

1 Introduction

Research in Human-Computer Interaction (HCI) demonstrates that the design of tools and procedures significantly impacts total human-system performance. Formal engineering models of cognitive and perceptual-motor processes can aid system design and evaluation. ACT-R is a formally specified theory of human cognition, perception, and action that enjoys wide scientific support and includes relatively rich

* Disclaimer: Any mention of commercial products or reference to commercial organizations is for information only; it does not imply recommendation or endorsement by the National Institute of Standards and Technology nor does it imply that the products mentioned are necessarily the best available for the purpose.

perception and motor action modeling capabilities [1] and [2]. However, ACT-R and other modeling frameworks have historically assumed the modeled user is seated at a desktop computer with a monitor, keyboard, and mouse. This assumption was both necessary and appropriate given the basic research from which said models were developed; decades of behavioral research in cognitive science and HCI were conducted in the traditional desktop computing environment. Due to recent advances in pervasive computing technology, this previously dominant interaction paradigm is rapidly giving way to mobile touchscreen devices [3].

Our project focuses on advancing modeling methods for multitouch tablet devices, motivated largely by their use in NIST's Biometric Web Services (BWS) project [4], which enables remote control of biometric devices by handheld touchscreen computers. In mission-critical systems, it is often difficult or impossible to gain access to the appropriate users in sufficient numbers and contexts. For example, it would be inadvisable—even dangerous—to deploy a large experimental biometrics system simply for research purposes at customs and border control. Instead, computational cognitive modeling can significantly augment existing usability testing methods by simulating human performance in these types of complex, dynamic systems. The current work will allow NIST's BWS project to model human operators using small handheld tablets and networked biometric sensors to capture biometric data, ultimately exploring and measuring effects of operator error, network delays, and sensor failure on total system performance. More importantly, the current work provides the basic functional foundation upon which to advance general modeling theory and practice across a variety of existing and emerging mobile task domains.

1.1 ACT-R

ACT-R is a computational cognitive architecture, a general theory of human cognition instantiated in an open-source modeling and simulation software package [2]. This means that it is a formally specified framework for constructing models of how people perform tasks, and these models generate quantitative predictions. ACT-R incorporates a motor planning and execution module adapted from another cognitive architecture, EPIC [5]. According to this model of motor planning and execution, cognition has a vocabulary of simple movement styles such as *ply* and other, more complex styles composed from those, such as to move the mouse cursor. Each movement is specified by features such as which hand, which finger(s), movement direction and movement distance. Movements are requested by central cognition and processed in stages by a motor module. However, ACT-R and other modeling frameworks still widely assume the modeled user is seated at a traditional desktop computer with a monitor, keyboard, and mouse. By extending the manual motor capabilities of ACT-R to include such new command vocabulary as *swipe*, *pinch*, and *rotate* gestures, we can better simulate the more complex user interactions that exist in the milieu of novel mobile touchscreen devices today.

2 ACT-Touch

ACT-Touch extends ACT-R's existing motor module by providing it with a simulated multitouch display device and multitouch display gesture motor vocabulary; ACT-Touch is implemented as Lisp code that is meant to load with ACT-R's software. ACT-Touch can be downloaded as a single archive from Cogscents, LLC's website [6]. The architectural component to ACT-Touch is a library of manual motor request extensions, including assumptions and changes specific to the multitouch task environment. One such difference between ACT-R and ACT-Touch is the default starting hand position. ACT-Touch assumes that in the mobile touchscreen domain, a user's hands no longer start at traditional home row positions (left and right index fingers positioned over the F and J keys respectively) because there is no longer a desktop keyboard present. Instead, a user's hands are assumed to start on both sides of the tablet (left hand on left side of tablet, right hand on right side of tablet).

The units of measurement for distance specifications in motor movement requests also differs between ACT-R (distance in "keys") and ACT-Touch (distance in pixels, at 72 ppi). Motor movement distance in ACT-R was measured in "keys" (the distance between two adjacent keys in the same row or the same column on the simulated desktop keyboard). This works well for modeling typing tasks in the desktop computing environment given consistency in key sizes and spacing for traditional QWERTY physical keyboards. However, in the newer mobile touchscreen computing environment, virtual keyboards can vary significantly. Virtual keyboard sizes vary across mobile devices based on physical differences in the maximum available touchscreen real estate. Virtual keyboards may even vary within a single device, depending on device orientation (landscape versus portrait mode) and whether the virtual keyboard is split; in addition to reducing key sizes, splitting the keyboard also changes the relative distance between some keys more so than others. For these reasons, ACT-Touch uses pixels to specify distances for motor movement requests.

ACT-Touch introduces a z-axis for motor movements, which represents vertical distance between the surface of the multitouch display device and the model's finger above it. ACT-Touch adds several basic motor movement styles (tap, swipe, pinch, and rotate gestures) that are commonly used across a variety of today's handheld mobile devices. As with ACT-R, ACT-Touch's basic movement styles are then combined to form more complex movements. The specific multitouch gestural commands currently implemented in ACT-Touch are: tap, peck-tap, peck-recoil-tap, tap-hold, tap-release, tap-drag-release, swipe, pinch, rotate, and move-hand-touch.

A tap gesture in ACT-Touch simulates the model's finger moving toward and momentarily contacting the surface of the multitouch display directly under the finger's current location; this is analogous to ACT-R's punch command for a traditional desktop keyboard, where the key to be pressed is already directly below the finger. A peck-tap gesture in ACT-Touch simulates moving the model's finger to a new location and tapping that location on the multitouch display (as a continuous movement). Peck-recoil-tap is similar, except the model's finger returns to its starting location after having tapped the display. Tap-hold simulates the model tapping and holding a finger on the surface of the multitouch display until a tap-release movement

is requested. Tap-drag-release simulates a drag-and-drop movement (e.g., the model will tap-hold the display surface under its finger, move said finger to a new location without breaking contact with the display surface, then release its finger from the display). For a swipe, the model moves the specified number of fingers (1-5, incrementing from index to pinkie and thumb) onto the display, moves them the specified distance and direction, then releases them from the display. For a pinch/reverse pinch gesture, the model moves the specified finger and thumb onto the display, moves them together/apart by the difference between the specified start- and end-widths (in pixels), then releases them from the display. A rotate gesture is similar (move finger and thumb to display, move them on display surface, release from display), but the distance between the model's digits remains constant as they are moved rotationally; direction of rotation is specified in radians.

The ACT-Touch distribution includes a simulated virtual multitouch display device (1,024 pixels wide x 768 pixels tall) with which a model can interact using the new motor movement commands described above. Default hand positions for the model are at either side of the display, centered approximately vertically. The ACT-Touch distribution also includes a virtual experiment window, a derivative of Mike Byrne's Experiment-Window ACT-R experiment instrumentation library [7]. The modified virtual experiment window allows modelers to build a multitouch display-based task environment and collect data for ACT-Touch. ACT-Touch, together with ACT-R, outputs a time-stamped series of predicted user behaviors. The instrumentation built into ACT-Touch supports capture of user latencies and errors within the simulated mobile touchscreen computing environment.

3 Model Validation

A critical next step is to compare ACT-Touch's motor movement predictions with human behavioral data. After testing and validating ACT-Touch's predictions for single-finger discrete tapping input (i.e., tap, peck-tap, and peck-recoil-tap), we will move on to testing predictions for more complex, continuous single-finger gestures, then progress in an orderly manner through two-, three-, and four-fingered multitouch gestures. As ACT-Touch predicts both movement latencies and their associated XY touch coordinates, we are collecting human data at a similarly fine level of granularity. These model validation efforts are currently under way, starting with implementation of customized touch-logging for tapping tasks on mobile handheld iOS devices (currently Apple iPads). We record a timestamped log of user touch events (XY screen coordinates) and corresponding system responses (e.g., button press detected, popover dismissed) via the actual mobile device with which they are completing the task.

3.1 Touch-Logging

We have already identified several touch-logging challenges of particular interest for model validation efforts; some pose immediate issues, while others may not apply

until the tasks we model become more complex and representative of real-world activities. Of immediate relevance is the question of sampling rates during scrolls: is there a single ideal sampling rate for these purposes or does it vary based on the complexity of the gesture and task being modeled? Sampling rates that are too low may not provide sufficiently detailed data for investigating more subtle effects in users' movement trajectories. On the other hand, higher sampling rates cause touch-logging files to become very large, very quickly, especially when system events and notifications are also logged.

Touch-logging for certain native iOS interface elements can be particularly nuanced. In native iOS applications with unusually small buttons (i.e., smaller than the minimum iPhone button size of 44x44 pixels recommended in Apple's Human Interface Guidelines [8]), the active touch area is automatically extended invisibly beyond the button. These invisible button extensions make the effective target size larger than the visible target, as is the case with the "Detail Disclosure" button or the standard "Back" button in navigation-based iPhone applications. Another example of enlarged hit areas occurs on the inner edges of a split iOS virtual iPad keyboard. Although it is fairly easy to log notifications of when the keyboard appears and disappears, capturing the actual user input event (i.e., timestamp and XY touch coordinates for a tap on the "hide keyboard button") that triggered the keyboard hide is not possible without additional custom code.

Finally, the native iOS magnifier loupe may also pose challenges for modeling certain tasks, as there are no system notifications to log when the magnifier loupe appears or disappears. Attempting to detect it by listening for long presses in a text field has not worked thus far, as the act of listening disables the loupe. Current options include adding invisible controls on top of the text field to detect the long press, or re-implementing the entire functionality of the magnifier loupe. It may be the case that the programming effort required to implement these types of iOS work-arounds would be better spent elsewhere. Additional work is necessary to compare and contrast relative touch-logging capabilities between different mobile development platforms and devices.

4 Future Work

Assumptions borne from directly adapting EPIC's model raise some interesting questions in the touchscreen domain. Fitts' Law requires a target width, but what exactly is the target of a swipe or a pinch? A long fast swipe used for scrolling quickly through multiple pages versus one intended to scroll to a more precise location within a single page? The short quick flick required to turn a virtual book page? Does it make more sense to construct two-, three-, and four-fingered swipes as different movement styles, as opposed to ACT-Touch's current implementation of a single movement style with a feature specifying the number of fingers?

The preceding questions pertain specifically to movement predictions for the gesturing hand, but what about the stabilizing hand? Recent anthropometric research suggests that people will cycle between different stabilizing hand positions when

completing extended tasks in an unsupported environment [9]. The authors suggest future work to examine how input gestures may alter the stabilizing hand [9], but we are more interested in the reverse question: how does the stabilizing hand impact input gestures? Furthermore, do different stabilizing postures facilitate or inhibit motor learning and fatigue for the gesturing hand? Although motor learning and fatigue have not been widely modeled in the ACT-R community to date, this may change as the body of HCI/HF literature in this area continues to grow.

5 Discussion

The current work, ACT-Touch, successfully augments the existing ACT-R cognitive architecture for modeling touch and gesture on mobile devices. Specific multitouch gestural commands currently implemented include tap, peck-tap, peck-recoil-tap, tap-hold, tap-release, tap-drag-release, swipe, pinch/reverse pinch, rotate, and move-hand-touch. A critical next step is to compare model predictions with basic behavioral data, starting with simple psychophysical tasks to examine variance between people and gesture types, and examine motor learning and manual fatigue. Among other research questions, what is the target of a swipe? A pinch? Fitts' Law requires a target width, and recent literature suggests that traditional application of Fitts' Law may not always be sufficient for 3-D gestures and small touchscreen smartphones [10], [11]. Future research will focus on understanding when and how traditional HCI methods need modification to accurately model users interacting with novel technologies. To do so, modeling tools themselves must continually evolve to incorporate new interaction paradigms. The current work provides a technical foundation for modeling the evolution of touch and gesture on novel mobile devices.

Acknowledgements. This work is funded in part by a Measurement Science and Engineering grant from the National Institute of Standards and Technology's Information Technology Laboratory (ITL) Grant Program. Federal Funding Opportunity 2012-NIST-MSE-01. Grant 60NANB12D134, "Formal Model of Human-System Performance," awarded to Cogscnt, LLC.

References

1. Anderson, J.R., Bothell, D., Byrne, M.D., Douglass, S., Lebiere, C., Qin, Y.: An integrated theory of the mind. *Psychological Review* 111(4), 1036–1060 (2004)
2. Carnegie Mellon University, <http://act-r.psy.cmu.edu/about/>
3. Dediu, H.: Apple sold more iOS devices in, than all the Macs it sold in 28 years (2012), <http://www.asymco.com/2012/02/16/ios-devices-in-2011-vs-macs-sold-it-in-28-years/> (retrieved)
4. Biometric Web Services, <http://bws.nist.gov/>
5. Kieras, D.E., Meyer, D.E.: An overview of the EPIC architecture for cognition and performance with application to human-computer interaction. *Human-Computer Interaction* 12(4), 391–438 (1997)

6. Cogscent, LLC, <http://cogscent.com/>
7. CHIL, <http://chil.rice.edu/projects/RPM/index.html>
8. Apple, iOS Human Interface Guidelines (2012), <http://developer.apple.com/library/ios/documentation/userexperience/conceptual/mobilehig/MobileHIG.pdf> (retrieved)
9. Feathers, D.J., Zhang, H.: Holding a Multi-touch Tablet with One Hand: 3D Modeling and Visualization of Hand and Wrist Postures. In: Proceedings of the HFES 56th Annual Meeting (2012)
10. Jo, J.H., Kim, I.: Adjusting Fitts' Paradigm for Small Touch-Sensitive Input Device with Large Group of Users. In: Proceedings of the HFES 56th Annual Meeting (2012)
11. Park, D.: Prediction of a Three-Dimensional Pointing Task through Extending the Motor Module of ACT-R. In: Proceedings of the HFES 56th Annual Meeting (2012)

A Page Navigation Technique for Overlooking Content in a Digital Magazine

Yuichiro Kinoshita, Masayuki Sugiyama, and Kentaro Go

Department of Computer Science and Engineering,
University of Yamanashi, Kofu, Yamanashi 400-8511, Japan
ykinoshita@yamanashi.ac.jp, masayuki.sugiyama@ttmuh.org,
go@yamanashi.ac.jp

Abstract. Although electronic book readers have become popular in recent years, page navigation techniques used for these readers are not necessarily appropriate for all kinds of books. In this study, an observation experiment is conducted to investigate how people read paper-based magazines. Based on the findings in the experiment, the authors propose new page navigation techniques specialized for digital magazines. The techniques adopt the operation of flipping through the pages. A user study confirms that the techniques are useful for overlooking content in a digital magazine and able to support readers to find articles that meet their interests.

Keywords: Digital book, electronic book reader, overlooking content, page navigation, turning pages.

1 Introduction

Electronic book readers or tablet devices, such as Amazon's Kindle [1] and Apple's iPad [2], have become popular in these years. With the increasing population of the users, the variety of digital books has also been increased. The varieties are generally divided into two types: independent books and magazines. In the cases of independent books, such as novels and comic books, readers usually start reading from the first page and turn over the pages one by one since they have a single consecutive story. On the other hand, magazines usually consist of many individual articles. Readers do not necessarily need to read from the first page. For this reading style, the conventional page navigation techniques may not be appropriate.

Several devices and techniques have been proposed to realize improved page navigation. Chen et al. [3] discussed navigation techniques for their dual-display electronic-book reader. Flipper [4] is a digital document navigation technique inspired by paper document flipping. TouchMark [5] is navigation techniques that use physical tabs to enable page thumbing and bookmarking. The techniques also preserved physical affordances of paper books. Meanwhile, Smart books [6] added context-awareness to electronic books. Although these devices and techniques achieved efficient page navigation, their applicability to magazine reading has not been discussed well. Also, most of the devices and systems are not specialized for digital magazine reading.

In this study, an experiment is conducted to observe how people read and interact with paper-based magazines. Based on the findings in the experiment, the authors propose new page navigation techniques specialized for digital magazines.

2 Observation Experiment of Magazine Reading

2.1 Methods

In order to find features of magazine reading, an observation experiment was conducted using a paper-based fashion magazine. For the purpose of comparison, the same experiment was also conducted for a novel and a comic book, which are usually with a single consecutive story. Ten university student in their twenties participated in the experiment. The participants sitting on a couch were asked to hold a magazine or book with their hands, as shown in Fig. 1, and read it for two minutes as their usual. Their reading behaviour was video recorded. After the reading, they were asked to respond to a questionnaire. An interview was also conducted based on their questionnaire responses.



Fig. 1. Observation experiment environment

2.2 Results

In the beginning of the magazine reading, eight out of ten participants looked at the table of contents. These participants also flipped through the pages. This implied that most participants tried to overlook the whole content in the magazine to find articles with their interests. For the novel and the comic book, all the participants started reading from the first page and did not flip through the pages.

During the experiment, nine participants found articles with their interests in the case of the magazine. To the question of ‘How did you find the pages with your interests?’, seven participants responded that they found them by flipping through the pages. Only one participant found them using the table of contents. This implied that the table of contents did not work for grasping articles in a magazine. In the cases of the novel and the comic book, four and nine participants found pages with their interests, respectively. Unlike the magazine, all the participants responded that they found the interesting pages by turning over the pages one by one. The result suggested that the way of reading magazines was different from that for novels or comic books.

In terms of the question of ‘How did you judge if the pages meet your interests or not?’ for the magazine, four participants responded ‘by looking at large headlines’

while other four responded ‘by picture images.’ On the other hand, in the cases of the novel and the comic book, all the participants who found pages with their interests responded that they judged a page after reading the beginning of the story. The process was also different between magazines and novels or comic books.

From these findings, we propose page navigation techniques to overlook content in a magazine with the operation of flipping through the pages.

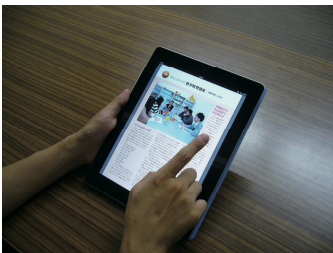
3 Proposed Page Navigation Techniques

3.1 Basic Page Navigation

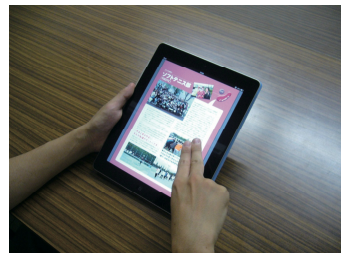
Likewise the conventional page navigation, the proposed techniques use tapping and swiping operations. Readers are able to turn over a single page by tapping on the left/right side of the screen or swiping on the centre of the screen using a single finger, as shown in Fig. 2(a). By tapping on the centre of the screen, a slider is shown on the bottom of the screen. The slider allows readers to roughly specify a page to which they want to move.

3.2 Continuous Multiple-Page Flipping

The results of the observation experiment showed that flipped through the pages enabled readers to overlook the whole content in the magazine. The proposed technique adopts the operation of flipping over multiple pages in addition to the aforementioned basic page navigation. By swiping with two fingers, as shown in Fig. 2(b), continuous multiple-page flipping starts. Users are able to stop flipping by tapping on the centre of the screen. The flipping speed is determined by the amount of swiping; long swiping makes the flipping faster and short swiping makes it slower.



(a) single page flipping



(b) continuous multiple-page flipping

Fig. 2. Page navigation in the proposed techniques

The page flipping speed is also changed by a weight assigned for specific pages. The flipping speed becomes slower when weighted pages are appeared. In the observation experiment, large headlines or picture images were keys to find pages with readers' interests. The weight was therefore assigned to pages with large headlines or images. The page weights increase readers' awareness to weighted pages and help them to see whether or not the pages meet their interests.

4 User Study

4.1 Methods

In order to investigate how the proposed techniques affect readers' page navigation, a user study was conducted for 12 participants. All of them were university students and familiar with electronic book readers. The study was conducted in the same environment used in the observation experiment. Three digital magazines were prepared from different categories: fashion, foods and gadgets. These magazines were presented to the participants using three types of applications running on Apple's iPad. Two applications were based on the proposed techniques, with and without the page weights. The other application was Apple's iBooks [7]. The combination of magazines and applications was switched per participants. The order of presentation was also balanced across participants.

Task 1. In the first task, the participants were asked to grasp the articles through a magazine in two minutes. After the reading each magazine, a paper-based task was conducted to examine how much content the participants have grasped. A paper form consisting of 45 article titles or headlines and 20 images was presented to the participants. Two-third of them were appeared in the magazine but the remaining ones were collected from other magazines. The participants selected all of recognizable headlines and images on the form.

Task 2. After Task 1, the participants were allowed to read magazines freely. This task observed the process to find pages that meet their interests and investigated how the proposed techniques affect their magazine reading on electronic book readers. The participants read each magazine for ten minutes. After each reading, the participants evaluated one of the applications through a questionnaire that used a five-point Likert scale and free form responses.

4.2 Results and Remarks

Task 1. In the reading with the proposed techniques, nine out of 12 participants used the continuous multiple-page flipping to overlook the whole articles. Some participants used the single page flipping in combination with the multiple-page flipping while other participants used the multiple-page flipping from beginning to end. In the case of iBooks, nine participants used the thumbnail function. For all the applications, three participants read the magazines only with the single page flipping and used neither the continuous multiple-page flipping nor the thumbnail function. The results of these participants were therefore excluded for the subsequent analyses.

Figure 3 shows the average number of pages displayed in two minutes. The participants checked much more pages when they used the proposed techniques. Significant differences were observed between the proposed techniques (with and without the page weights) and iBooks ($p < 0.01$). These results demonstrated that the proposed techniques have sufficient performance for overlooking the articles through a magazine. Figures 4 shows the average numbers of recognized headlines and images,

correctly selected on the form. In terms of the recognized headlines, no significant difference was observed between the applications. However, in the case of images, the participants recognized more images in the cases of proposed techniques. Significant differences were observed between the proposed techniques (with and without the page weights) and iBooks ($p < 0.05$). This implied the techniques were especially useful for magazines consisting of many picture images.

Task 2. Depending on the applications used in the task, different processes were observed to find pages with readers' interests. In the cases of the proposed techniques, most of the participants used the continuous multiple-page flipping. When they found something interesting, they stopped the flipping by tapping on the screen and checked the previous and/or next pages using the single page flipping. After reading these pages, they started the multiple-page flipping again. When the page flipping was reached to the end of the magazine, they used the slider and moved back to the first page. In the case of iBooks, most of the participants used the thumbnail function and

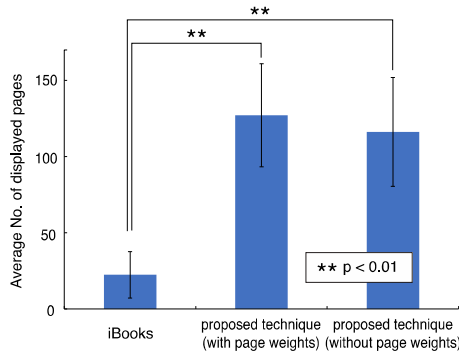


Fig. 3. Average number of pages displayed in two minutes

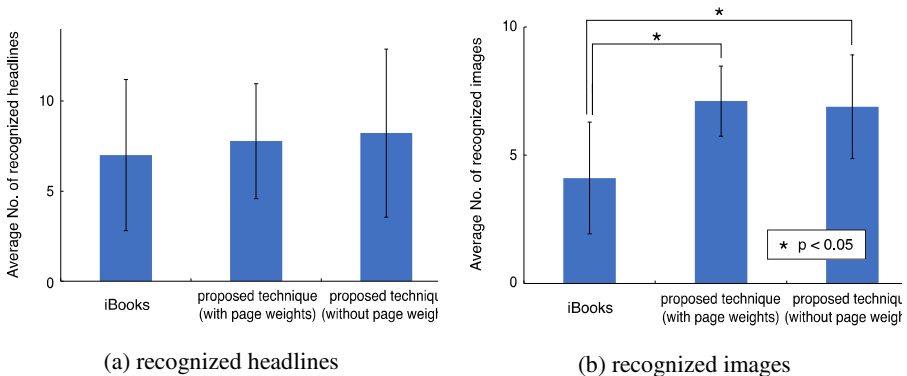


Fig. 4. Average numbers of recognized headlines and images

selected a page from the thumbnails. After reading the page, they moved back to the thumbnail function and selected another page. They repeated these operations to find pages with their interests. In this case, the number of times they used the single page flipping was fewer than that observed in the cases of the proposed techniques.

In terms of the questionnaire, the average evaluation score to the statement of ‘The operation of was similar to paper-based books/magazines.’ was 3.6 for the proposed techniques while the score was 2.2 for iBooks. Here ‘1’ and ‘5’ correspond to strongly disagree and strongly agree, respectively. The proposed techniques provided paper-like operation and enabled users to flip through the pages for overlooking content in a digital magazine.

5 Conclusion

This study first conducted the observation experiment to investigate features of paper-based magazine reading. The result of the experiment led the authors to the following findings. First, most readers try to overlook the whole content in a magazine. Second, readers find articles with their interests by flipping through the pages. Third, large headlines or picture images are keys to find articles with readers’ interests. Based on these findings, a page navigation technique for digital magazines with the operation of continuous multiple-page flipping were proposed. The user study confirmed that the proposed techniques were valid for overlooking content in a digital magazine. The techniques will help users to find articles with their interests more easily.

Future studies will address the investigation of appropriate page flipping speed according to the page weights as well as the discussion about another approach to overlook content in digital magazines.

References

1. Amazon: Kindle, <http://www.amazon.com/b?node=133141011>
2. Apple: iPad, <https://www.apple.com/ipad/>
3. Chen, N., Guimbretiere, F., Dixon, M., Lewis, C., Agrawala, M.: Navigation techniques for dual-display e-book readers. In: Proceedings of ACM SIGCHI Conference on Human Factors in Computing Systems, pp. 1779–1788 (2008)
4. Sun, L., Guimbretière, F.: Flipper: a new method of digital document navigation. In: ACM CHI 2005 Extended Abstracts on Human Factors in Computing Systems, pp. 2001–2004 (2005)
5. Wightman, D., Ginn, T., Vertegaal, R.: TouchMark: flexible document navigation and bookmarking techniques for e-book readers. In: Proceedings of Graphics Interface 2010, pp. 241–244 (2010)
6. Beer, W., Wagner, A.: Smart books: adding context-awareness and interaction to electronic books. In: Proceedings of the 9th International Conference on Advances in Mobile Computing and Multimedia, pp. 218–222 (2011)
7. iBooks, <http://www.apple.com/apps/ibooks/>

Effect of Unresponsive Time for User's Touch Action of Selecting an Icon on the Video Mirror Interface

Kazuyoshi Murata¹, Masatsugu Hattori², and Yu Shibuya¹

¹ Kyoto Institute of Technology, Kyoto, Japan
{kmurata, shibuya}@kit.ac.jp

² Kansai Ohkura Senior High School, Osaka, Japan
hattori@hi.cis.kit.ac.jp

Abstract. Contactless input methods implementing body motion allow users to control computer systems easily and enjoyably. We focus on the “video mirror interface” as an example of these methods. A user of the video mirror interface can operate the computer system by selecting virtual objects on a screen with his/her hand. However, if a selection operation is completed as soon as the user touches the virtual object, erroneous selections will frequently occur. Therefore, it is necessary to insert a certain period of unresponsiveness after a user's touch action to prevent selection error. We evaluate effects of the unresponsive time when selecting a virtual object in a video mirror interface. The result of an experimental evaluation indicates that an acceptable range for the unresponsive time is 0.3 to 0.5 s.

Keywords: Video mirror interface, unresponsive time, touch action.

1 Introduction

With the advances in computer technology, contactless input methods implementing body motion have been studied. Such input methods allow users to control computer systems easily and enjoyably. For example, Shibuya et al. [1] proposed the concept of “Action Interface” and provided practical examples of its application. The user of these applications can control a computer system to play a virtual instrument or a game using his/her body action. Hosoya et al. [2] proposed the “Mirror Metaphor Interaction System” to physically interact with CG icons or objects using touch-based actions. In addition, similar systems have been proposed for practicing sports [3][4]. These systems capture a video image of the user's body and display a mirrored video image along with some virtual objects superimposed on a screen in front of the user. The user can operate the computer system by using various body motions (e.g. manipulating the virtual objects with his/her hand). We call such systems a “video mirror interface” (an example of which is shown in Fig. 1).

In a video mirror interface, if a selection operation is completed as soon as the user touches the virtual object, erroneous selections will frequently occur. This is because the user cannot select the target object without interference from other objects that are

tightly-spaced and closely proximate to the target object. Therefore, it is necessary to insert a certain period of unresponsiveness after a user's touch action to prevent selection error.

In this paper, we focus on the effects of the unresponsive time when selecting a virtual object on a video mirror interface.

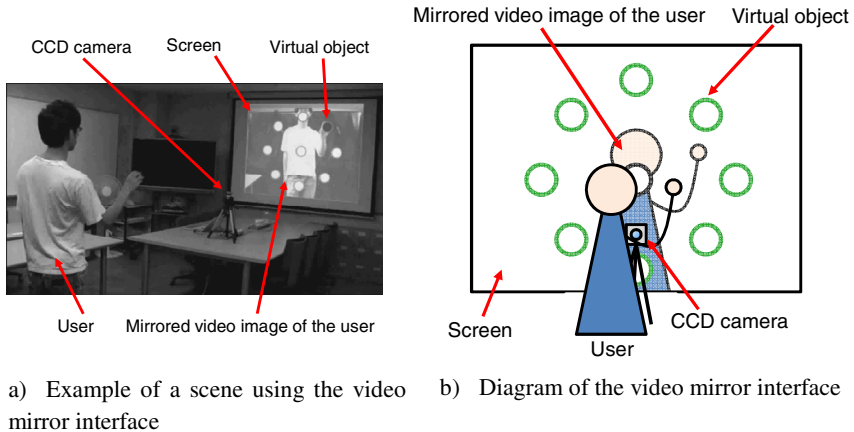


Fig. 1. Example of the video mirror interface

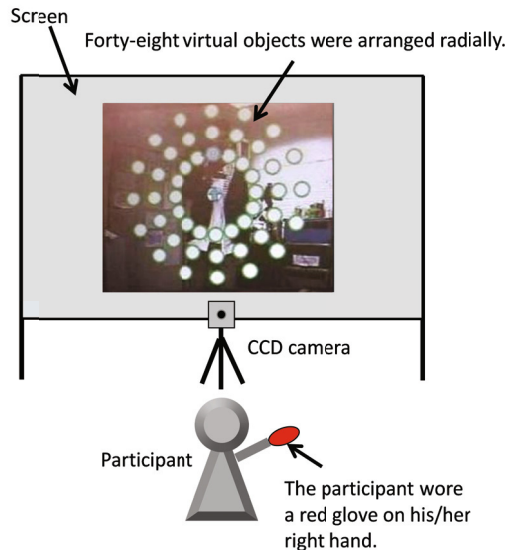


Fig. 2. Experimental system configuration

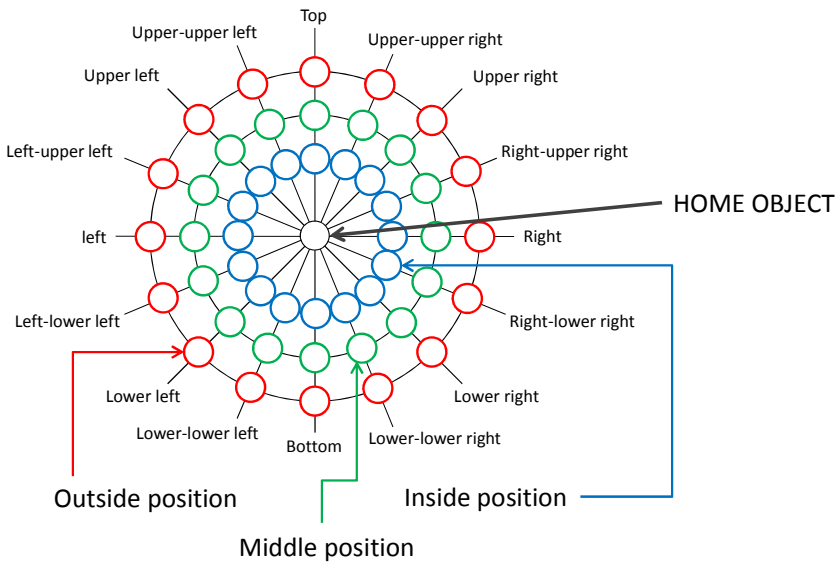


Fig. 3. Alignment of virtual objects

2 Experiment

An experimental evaluation was conducted to confirm the relationship between the duration of unresponsive time and the erroneous selection of virtual objects. Fig.2 shows the configuration of the experimental system. The system consisted of a CCD camera, a desktop PC, a projector and a screen. A user's life-sized video image was captured from the CCD camera putted in front of the user. This captured video image was sent to the PC and the PC transformed it to a mirrored video image. Simultaneously virtual objects generated by the PC and they are overlapped on the mirrored video image. Finally, the processed video image was projected on the screen set up in front of the user.

Fig.3 shows an alignment of virtual objects. Forty-eight virtual objects were arranged on each circumference of a circle with radius 65, 100, and 135 px, and one virtual object was positioned at the center of the other objects. The diameter of a virtual object was 24 px.

The participants were asked to wear a red glove on their right hands; this glove was used to detect the position of the hand by using image recognition. In this experiment, we used a glove to enhance the accuracy of detection in this system. However, contactless sensing devices, such as Kinect [5], is begun to launch recently. If these contactless devices are grown in performance, the video mirror interface can be implemented by using these contactless devices in the future.

3 Procedure and Experimental Design

In this experiment, participants were asked to perform the task described below with nine different durations of unresponsive time (0.0, 0.1, 0.2, 0.3, 0.5, 0.7, 1.0, 1.5, and 2.0 s). Below, we describe the procedure undertaken by each participant.

1. The participant was asked to stand in front of the screen.
2. The participant was then asked to select the virtual object positioned at the center of radially arranged virtual objects (hereafter the center object is referred to as the "HOME OBJECT"). As soon as the participant selected the HOME OBJECT, the appearance of the virtual object which was randomly selected was changed, hereafter referred to as the "TARGET OBJECT."
3. Next, the participant was asked to select the TARGET OBJECT. If he/she selected an incorrect virtual object, it was referred to as a "selection error." When such a selection error occurred, the participant repeated the selection task until he/she successfully selected the TARGET OBJECT.
4. The participant was asked to repeat the selection task (step 2 to 3) 48 times.

Ten volunteers were recruited from our university to participate in the experiment. We used a within-subject design in the experiment. The independent variables were the duration of the unresponsive time and the alignment of virtual objects (Fig. 3). Dependent measures included the number of selection errors per one selection operation. In addition, participants were asked to answer a questionnaire for subjective evaluation.

4 Result and Discussions

Fig. 4 shows the result of the number of selection errors per one selecting operation. We carried out an analysis of variance and there were main effect on the duration of unresponsive time ($F_{(5,45)}=713.255$, $p<.01$) and the alignment of virtual objects ($F_{(2,18)}=189.347$, $p<.01$). A post-hoc test showed that there were significant differences between every combination of the duration of the unresponsive time except 0.3 and 0.5, 0.3 and 0.7, and 0.5 and 0.7 s. In addition, there were significant differences between every combination of the alignment of virtual objects.

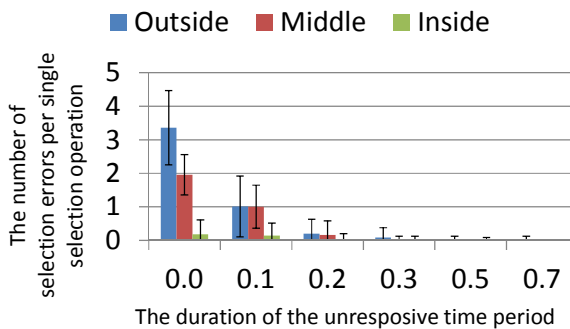


Fig. 4. Number of selection errors

When the duration of the unresponsive time was 0.0 s, multiple selection errors occurred in the case of the TARGET OBJECT positioned at the outside and middle position. In addition, the selection errors were not avoided even when the duration of the unresponsive time was 0.1 s.

If the TARGET OBJECT was positioned at the outside and middle position, the participants needed to pass over objects positioned at the inside position. Therefore, selection errors were not avoided if the duration of the unresponsive time was 0.2 s or shorter. On the other hand, if the duration of the unresponsive time was 0.3 s or longer, there were few selection errors.

As for the number of selection errors at the inside position, the number of selection errors was 0.03 when the duration of the unresponsive time was 0.2 s, and that was less than 0.01 when the duration of the unresponsive time was 0.3 s or longer. If the TARGET OBJECT was closely proximate to other objects such as an object positioned at inside position, it was difficult for participants to select the TARGET OBJECT without interference from other objects. However, if an unresponsive time was 0.3 s or longer, erroneous selection was able to avoid.

For subjective evaluation, we also asked the participants to answer a questionnaire. For example, the question “Did you have an unpleasant feeling?” and “Were you fatigued with the selection task?” were asked after he/she completed each task. Fig.5 shows the results of the question: “Did you have an unpleasant feeling?” A Friedman test revealed a significant main effect on the duration of the unresponsive time ($\chi^2=48.936$, $p<.01$). A post-hoc test revealed that the score of 0.0 s unresponsive time was significantly higher than that of 0.2 and 0.3 s unresponsive time. Moreover, the score of 1.0 and 1.5 s unresponsive time were higher than that of 0.5 s unresponsive time. In addition, when the duration of the unresponsive time was 0.0 or 0.1 s, some participants said that the duration of the unresponsive time was too short to select the target object without selection error. On the other hand, when the duration of the unresponsive time was greater than 0.7 s, some participants said that this was too long.

Fig.6 shows the results of the question: “Were you fatigued with the selection task?” A Friedman test revealed a significant main effect on the duration of the unresponsive time ($\chi^2=49.835$, $p<.01$). A post-hoc test revealed that the score of 1.0 and

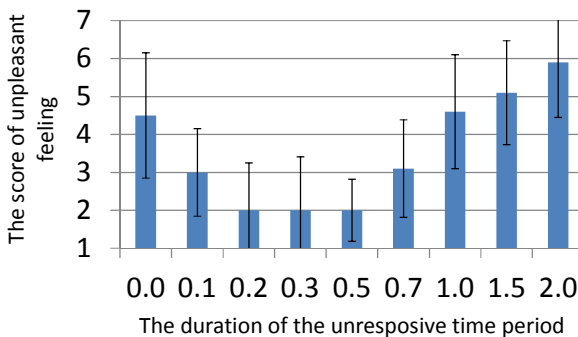


Fig. 5. Results of the question: “Did you have an unpleasant feeling?”

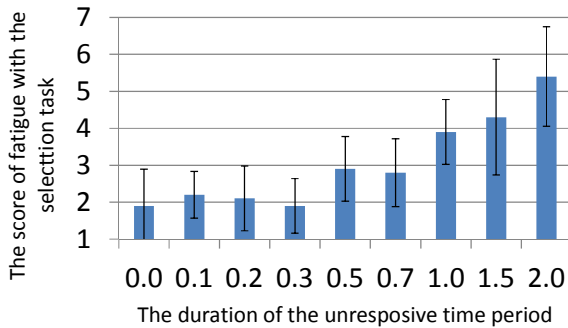


Fig. 6. Results of the question: “Were you fatigued with the selection task?”

1.5 s unresponsive time were higher than that of 0.3 s unresponsive time. The score of 2.0 s unresponsive time was also significantly higher than that of 1.0 s or less unresponsive time. In addition, some participant said that maintaining their hand in that position for that long was tiring when the duration of the unresponsive time was 1.0 s or longer.

These experimental results indicate that an unresponsive time which is 0.2 s or shorter increases selection errors. Moreover, participants feel unpleasant if the unresponsive time is too short. On the other hand, if the unresponsive time is 0.7 s or longer, participants tire while keeping their hands in the air. These results indicate that, for few erroneous selections and high participant evaluation results, an acceptable range for the unresponsive time is 0.3 to 0.5 s.

5 Conclusion

In this paper, we focus on the effects of the unresponsive time when selecting a virtual object on a video mirror interface. An experimental evaluation was conducted and the result indicated that erroneous selections were able to avoid if the duration of the unresponsive time is 0.3 s or longer. In addition, when the duration of the unresponsive time was 0.7 s or longer, participants said that the duration of the unresponsive time is too long and that maintaining their hand in that position for that long was tiring. For few erroneous selections and high participant evaluation results, these results indicated that an acceptable range for the unresponsive time is 0.3 to 0.5 s.

References

1. Shibuya, Y., Takahashi, K., Tamura, H.: Introduction of Action Interface and Its Experimental Uses for Device Control, Performing Art, and Menu Selection. In: The 7th IFAC Symposium on Man-Machine Systems, pp. 71–76 (1998)

2. Hosoya, E., Kitabata, M., Sato, H., Harada, I., Nojima, H., Morisawa, F., Mutoh, S., Onozawa, A.: A Mirror Metaphor Interaction System: Touching Remote Real Objects in an Augmented Reality Environment. In: The Second IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR 2003), pp. 350–351 (2003)
3. Hämäläinen, P.: Interactive Video Mirrors for Sports Training. In: The 3rd Nordic Conference on Human-Computer Interaction, pp. 199–202 (2004)
4. Kuramoto, I., Inagaki, Y., Shibuya, Y., Tsujino, Y.: Augmented Practice Mirror: A Self-learning Support System of Physical Motion with Real-Time Comparison to Teacher's Model. In: Duffy, V.G. (ed.) ICDHM 2009. LNCS, vol. 5620, pp. 123–131. Springer, Heidelberg (2009)
5. Microsoft Kinect for Xbox 360, <http://www.xbox.com/en-US/kinect>

Evaluation of a Soft-Surfaced Multi-touch Interface

Anna Noguchi, Toshifumi Kurosawa, Ayaka Suzuki, Yuichiro Sakamoto,
Tatsuhito Oe, Takuto Yoshikawa, Buntarou Shizuki, and Jiro Tanaka

University of Tsukuba, Japan

{noguchi,kurosawa,ayaka,sakamoto,tatsuhito,
yoshikawa,shizuki,jiro}@iplab.cs.tsukuba.ac.jp

Abstract. “WrinkleSurface”, which we developed by attaching a gel sheet to a FTIR-based touchscreen, enables a user to perform novel touch motions such as Push, Thrust, and Twist_CW (clockwise), and Twist_CCW (counterclockwise). Our research is focused on the evaluation of this soft-surfaced multi-touch interface. Specifically, to examine how a user can input our novel input methods precisely, we evaluated the user’s performance of each method by two to nine levels of target acquisition task. As a result, we found some points to be improved in our recognition algorithm in order to increase the success rate of Push and Thrust. In addition, a user can input Twist before the level of six because the success rate of Twist was high up to that level.

Keywords: Touchscreen, tabletop, haptic interface, FTIR, tangential force sensing, pressure sensing.

1 Introduction

In conventional touchscreen interaction, input is limited to the coordinates of human fingers’ contact areas. Recently, many researchers have worked on novel input that exceeds these coordinates to enrich touchscreen interaction [1, 2, 5, 6, 8, 10, 12]. We also developed “WrinkleSurface” to explore a wide variety of inputs in touchscreen interaction [9]. WrinkleSurface is a soft-surfaced touchscreen that enables a user to perform novel touch motions such as pushing, thrusting, and twisting (Fig. 1), in addition to conventional motions like drag and pinch. We named these input methods Push, Thrust, and Twist and presented some applications in touchscreen interaction. It is also possible to detect the strength and direction of the motion from the wrinkles caused by the motion.

Our research is focused on the evaluation of a soft-surfaced multi-touch interface. Specifically, to examine how a user can input our novel input methods precisely, we evaluated the user’s performance of each method.

We begin by describing the previous works on input methods that achieved a variety of input in multi touch interfaces, and soft-surfaced touch interfaces. Next, we present the hardware design and recognition methods of our soft-surfaced touchscreen. We end with a user study that shows user’s performance of our novel input methods using WrinkleSurface.

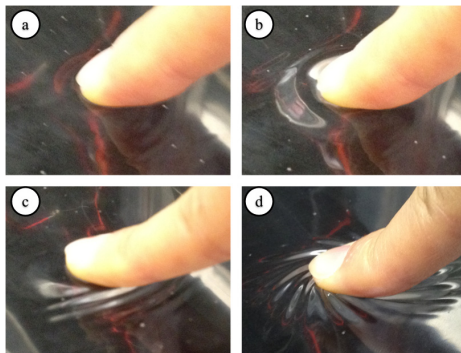


Fig. 1. a) Touch, b) Push, c) Thrust, and d) Twist

2 Related Work

Our research is focused on the evaluation of a soft-surfaced multi-touch interface. There are many works related to the research of this paper that have developed various of inputs in multi touch and soft-surfaced touch interfaces. However, few researchers have evaluated of their methods.

2.1 Input Methods in Touch Interaction

Some researches have attempted to obtain information other than coordinates by utilizing the shape of finger’s contact area on the touchscreen surface in order to enrich interaction. Wang et al. and Dang et al. focused on oblique touch input and developed a way of detecting the finger orientation [1, 2, 12]. Our WrinkleSurface can sense the finger orientation as well, but our system is mainly intended not to utilize the information concomitant with finger’s contact, but to develop novel input methods such as pushing, thrusting, and twisting.

Some have researched recognizing the finger posture above the touchscreen. Takeoka et al. proposed Z-touch, which utilizes the slant and direction of fingers in touchscreen interaction [10]. Z-touch uses infrared laser plane and a high-speed camera to recognize the finger posture in the space above the touchscreen. Our research is different from this because it is based on the force sense feedback deriving from direct contact with the input surface.

Heo et al. and Lee et al. enabled detection of horizontal movement on the surface of a device and recognized various types of inputs [6, 8]. They made a cover with sensors at the bottom and the side frame, enclosed a device within it, and detected the horizontal movement. Harrison et al. proposed Shear as a novel input that utilizes a tangential force to a screen’s surface [5]. Shear is “a supplemental analog 2D input channel” and can be used with conventional touch input. Our WrinkleSurface detects the force through the deformation of the gel-sheet (i.e., wrinkles).

Wang et al., Dang et al., and Lee et al. evaluated their approaches on the ordinary touchscreens, but we focused on the evaluation of a soft-surfaced touchscreen.

2.2 Soft-Surface Touch Interactions

Vlack et al. proposed GelForce, which detects strength and direction of forces applied to the surface of an elastic material [11]. It consists of a CCD camera and two layers of colored markers embedded in a transparent silicone rubber. This research developed soft-surface touch interactions. In contrast, our research focused not only the development but also evaluation of soft-surface touch interaction.

Takei et al. proposed a tabletop tangible interface, ForceTile [7]. The tile interface consists of an elastic body and markers. Cameras and infrared transmitters are placed underneath the tabletop, and they sense the position, rotation, and ID of interface and calculate the force vector of deformation. WrinkleSurface recognizes the input strength without any markers. Therefore, rear projection is possible, which is an advantage for a touchscreen. Our novel input Push, Thrust, and Twist are recognized by the wrinkles caused on the touchscreen, which is a novel recognition method.

Sato et al. showed PhotoelasticTouch, which recognized deformation of transparent elastic material as an input without using visual markers. It is made from transparent elastic material and consists of an LCD and an overhead camera both fitted with a quarter-wavelength filter. When force was applied to the elastic material, the deformed area transforms incoming light into elliptically polarized light, which is captured by the camera. Position and size of the deformed area and direction of the force can be calculated, and interactions such as pinching, pushing, and pulling became recognizable. Its weakness was the positioning of the camera. It was placed above the touchscreen and sometimes users' body parts (e.g., head) interfered with capturing hand images. In the case of WrinkleSurface, the IR camera is placed under the panel, and we can utilize the wrinkles caused on the touchscreen without such interruptions. In addition, WrinkleSurface enabled us to recognize novel input (thrusting and twisting) without any occlusions.

Fukumoto attached a soft-gel based transparent film named "PuyoSheet" onto the surface of the touchscreen. PuyoSheet, combined with soft-gel based small dots "PuyoDots", provided a button-push feeling to the fingertips [3]. WrinkleSurface provides novel input utilizing the softness of its surface in addition to tactile feedback by the restitution of the soft-gel surface. Fukumoto evaluated the performance and impressions of PuyoSheet and PuyoDots with a handheld device, but WrinkleSurface is a multi-touch tabletop interface.

3 WrinkleSurface

WrinkleSurface is a touchscreen based on frustrated total internal reflection (FTIR) [4]. A transparent urethane soft gel sheet (hereinafter gel sheet) about

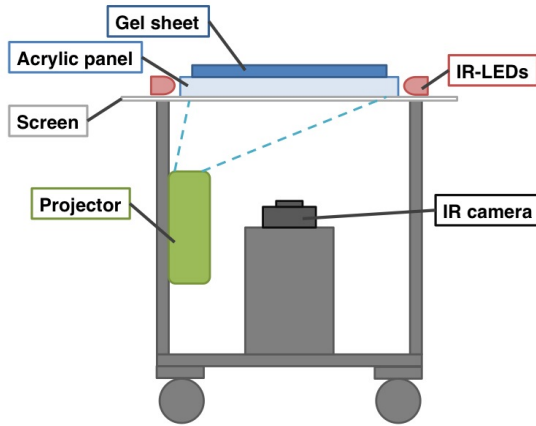


Fig. 2. Hardware setup of WrinkleSurface

3.5 mm thick is attached onto the surface of an acrylic panel (Fig. 2). Each touch motion, such as pushing, thrusting, and twisting, makes characteristic wrinkles on the surface. Utilizing this, we have extracted the features of these wrinkles to detect three novel input methods. Moreover, the strength of each input can be detected. Fig. 1 shows the ordinary Touch and our input methods: Push, Thrust, and Twist. When a user pushes vertically into the panel with a certain strength, it results in Push (Fig. 1b). Wrinkles do not appear on the gel sheet in both Touch and Push. When a user slides the finger while Pushing, wrinkles appear in the direction in which a user slides the finger (Fig. 1c). When a user rotates the finger while Pushing, wrinkles appear around the finger (Fig. 1d).

3.1 Hardware Setup

As shown in Fig. 2, WrinkleSurface consists of an acrylic panel attached to a gel sheet, 28 infrared LEDs, an infrared camera, a projector, and a screen for projection. WrinkleSurface is placed 1100 mm above floor level. The acrylic panel is $590 \times 450 \times 10$ mm, and the gel sheet is $500 \times 400 \times 3.5$ mm. 14 IR LEDs (OPTOELECTRONICS CO., LTD., SFH4550) are attached lengthways along the acrylic panel. We also placed the IR camera (Point Grey Research, Dragonfly2) and the projector underneath the panel. The bottom of the acrylic panel and the camera lens are 330 mm apart. Under the acrylic panel, we put a tracing paper of 40 g/m^2 paper density as a screen for projection.

In this system, we used FTIR to detect input. After attaching the gel sheet to the FTIR touchscreen, we checked and confirmed that the FTIR mechanism worked properly. FTIR-based touchscreen and WrinkleSurface differ in diffuse reflection. In the case of WrinkleSurface, the diffuse reflection takes place not only in the finger contact area like for FTIR-based touchscreen but also in the

wrinkled area. By capturing the diffuse reflection image with an IR camera, we can obtain the shape and angles of the wrinkles appearing on the gel sheet.

3.2 Recognition Techniques

In this section, we describe the recognition technique of our novel inputs and the strength of these inputs. WrinkleSurface recognizes each input by means of the following process.

1. Extracting three characteristic parameters from the image processing: “roundness”, “magnitude of the wrinkle vector”, and “rotation degree”.
2. Defining the likelihood function associated with each input and the characteristic parameters.

“Roundness” is the roundness of the combined area consisting of the finger’s contact area and the wrinkled area. “Magnitude of the wrinkle vector” is expressed by the Euclidean distance between the gravity centers of the finger’s contact area and the wrinkled area. “Rotation degree” is obtained by comparing the inter frame differences of the finger direction calculated using the algorithm developed by Wang et al [12]. “Rotation degree” is only used for the Twist. To recognize three input methods, we experimentally developed a likelihood function that uses these characteristics as its parameters. The system also recognized the strength of each input from the parameters listed in Table. 1.

Table 1. Range of parameters for each input method

Input	parameter	range
Push	luminance value	50-110
Thrust	moving distance	0-60 (pixel)
Twist_CW	rotation angle	20-80 (degree)
Twist_CCW	rotation angle	20-80 (degree)

4 Applications

We developed three applications taking advantage of the features of WrinkleSurface. WrinkleGeo edits the geographical terrain utilizing wrinkles that appear on WrinkleSurface. WrinkleMesh distorts the image into the spiral pattern, utilizing the Twist. WrinkleIcon operates icons making good use of the repulsion of the gel sheet. Using WrinkleIcon, a user can flick out or gather icons.

5 Evaluation

To examine how a user can perform input precisely using our novel input methods, we evaluated the user’s resolution of each method’s parameter such as

the strength of Push, the moving distance of Thrust, and the rotation angle of Twist_CW (Twist clockwise) and Twist_CCW (Twist counterclockwise). To simplify the experimental setup, we did not use the projector or the screen for projection in this experiment (Fig. 3) and covered the frame of WrinkleSurface with a blackout curtain in order to block out sunlight from the camera (Fig. 4).

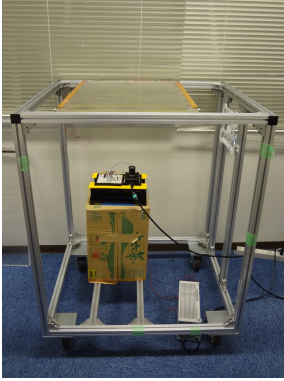


Fig. 3. Experiment environment

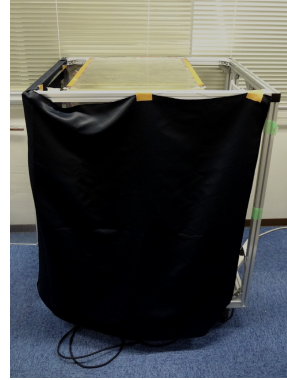


Fig. 4. WrinkleSurface surrounded by the blackout curtain

5.1 Participants

We recruited seven male and five female participants, aged 21-24, who were university (undergraduate and graduate) students majoring in computer science. All 12 participants were right-handed, and we asked them to use only their right index finger to complete the task to make the experiment conditions identical.

5.2 Task

Each participant engaged in a target acquisition task (Fig. 5). They were asked to adjust the size of the yellow circle (i.e., cursor) by controlling the parameter of an input method between two green circles (i.e., target). The cursor and the target were shown on the display. When a participant kept the cursor within the target for 1000 ms, the trial was a success. When 5000 ms elapsed after a participant had started a trial, the trial was a failure.

For each method, the range of the parameter was divided into two to nine levels and represented as a target (Fig. 6). Therefore, the experimental application provided 44 types of targets ($2+3+\dots+9$) for each parameter. For each target, the inner and outer circle of the target represented the minimum and maximum values of the target, respectively (e.g., 50 and 80 when the range of Push was divided into two levels). In this experiment, six sets of trials were given for each target. Thus, each participant had 1056 trials:

$$\begin{aligned}
 & \textit{Target type} : 44 \\
 & \times \textit{Set} : 6 \\
 & \times \textit{Input method} : 4 \\
 & = 1056.
 \end{aligned}$$

In total, this experiment had 12672 trials.

It took about three hours for each participant to complete the task. The experiment was conducted in a casual atmosphere. The participants could rest between the trials and talk freely to the other members including the experimenter in the laboratory.

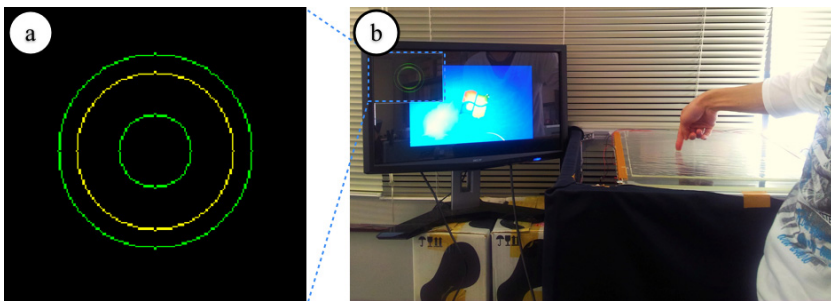


Fig. 5. a) Screen used in the task with cursor (yellow circle) and target (two green circles). b) Participant operating WrinkleSurface.

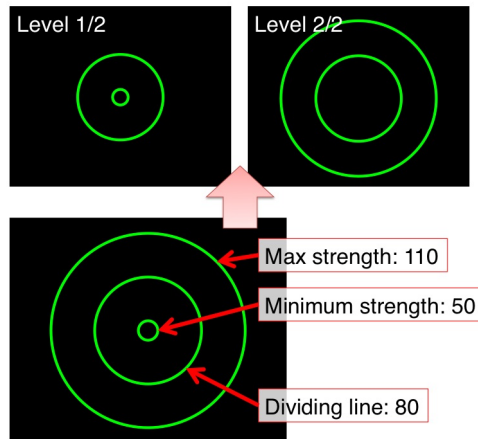


Fig. 6. Two targets when the range of Push was divided into two levels

6 Result and Discussion

Fig. 7 shows the average success rate of each method. Due to a system error, we could not use one participant's results. Thus, we had to calculate the average from the other 11 participants' results. The grand average was 27.2% in Push, 57.9% in Thrust, 49.9% in Twist_CW, and 50.5% in Twist_CCW.

From these results, we also identified some points to be improved in our recognition algorithm. One drawback of our algorithm is that the overall success rate of Push was too low. This was because when the target was small, the luminance value was high enough, but other elements caused the failure. We found out that WrinkleSurface was mistaking Push for Thrust when a participant pushed the gel sheet strongly. To solve this failure, we could stop distinguishing Push from Thrust or only recognize strong Thrust as Thrust.

The success rate of Thrust was higher than those of the other three inputs in the levels over four. However, the error rate of Thrust increased as the level increased. By examining of the error of Thrust, we found that there were two causes. First, WrinkleSurface had mistaken Thrust as Push or Twist because the movement of wrinkle vector is small. Second, the system changed the recognition into Touch when the force applied to WrinkleSurface was weakening while a user was performing Thrust. Both causes occurred when a participant was moving his/her fingers. Moreover, in the case of Thrust, the success rate of females was lower than that of males. This is because women are physically weaker than men in general. From the experimental observation and participants' comment, it seemed that it was difficult for female participants to keep their initial strength of Push to the end of the movement. Both male and female participants also commented that they had pain in their fingers, especially doing Thrust. To solve these problems, we believe that the elapsed time of thrusting is a possible parameter to improve the accuracy of Thrust recognition.

As shown in Fig. 7, the success rates of Twist_CW and Twist_CCW did not differ much. Both varied inversely to the levels except the sixth. The major cause

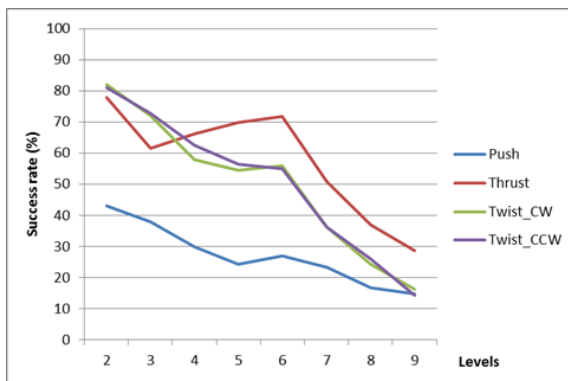


Fig. 7. Average success rate of each input

of failure of Twist was due to the trembling of the rotation angle; we observed that a user found it difficult to keep the target angle stable. This result indicates that Twist (both CW and CCW) is suitable for continuous input, or a use in rough situation.

Some participants made positive statements about WrinkleSurface. Many said that they would prefer a much softer surface. The softness depends on gel sheets, so we plan to experiment with another softer gel sheet. Some participants suggested CG modeling and paint application for WrinkleSurface. We will continue to evaluate and improve both the software and hardware of WrinkleSurface.

7 Conclusion

Our research is focused on the evaluation of a soft-surface multi-touch interface. To this end, we presented an evaluation of “WrinkleSurface”, which we developed by attaching a gel sheet to a FTIR-based touchscreen. In the evaluation, we obtained the grand average of 27.2% in Push, 57.9% in Thrust, 49.9% in Twist_CW, and 50.5% in Twist_CCW, and positive statements for WrinkleSurface from some participants. From these results, we found some points to be improved in our recognition algorithm in order to increase the success rate of Push and Thrust. The results also suggest that a user can input Twist before the level of six. We plan to continue to evaluate and improve both the software and hardware of WrinkleSurface.

Acknowledgements. First and foremost, the authors are grateful to Dr. Masaaki Fukumoto, Executive Research Engineer at NTT DOCOMO’s Frontier Technology Research Group. He provided us PuyoSheet, which is used in WrinkleSurface, and gave us advice. We also thank Prof. Simona Vasilache and Ms. Audrey A. M. Portron for their helpful comments on this paper.

References

1. Dang, C.T., André, E.: Usage and recognition of finger orientation for multi-touch tabletop interaction. In: Campos, P., Graham, N., Jorge, J., Nunes, N., Palanque, P., Winckler, M. (eds.) INTERACT 2011, Part III. LNCS, vol. 6948, pp. 409–426. Springer, Heidelberg (2011)
2. Dang, C.T., Straub, M., André, E.: Hand distinction for multi-touch tabletop interaction. In: Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces, ITS 2009, pp. 101–108 (2009)
3. Fukumoto, M.: Puyosheet and puyodots: simple techniques for adding “button-push” feeling to touch panels. In: CHI 2009 Extended Abstracts on Human Factors in Computing Systems, CHI EA 2009, pp. 3925–3930 (2009)
4. 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, UIST 2005, pp. 115–118 (2005)

5. Harrison, C., Hudson, S.: Using shear as a supplemental two-dimensional input channel for rich touchscreen interaction. In: Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems, CHI 2012, pp. 3149–3152 (2012)
6. Heo, S., Lee, G.: Force gestures: augmented touch screen gestures using normal and tangential force. In: Proceedings of the 2011 Annual Conference Extended Abstracts on Human Factors in Computing Systems, CHI EA 2011, pp. 1909–1914 (2011)
7. Kakehi, Y., Jo, K., Sato, K., Minamizawa, K., Nii, H., Kawakami, N., Naemura, T., Tachi, S.: Forcetile: tabletop tangible interface with vision-based force distribution sensing. In: ACM SIGGRAPH 2008 New Tech Demos, SIGGRAPH 2008, p. 17:1 (2008)
8. Lee, B., Lee, H., Lim, S.-C., Lee, H., Han, S., Park, J.: Evaluation of human tangential force input performance. In: Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems, CHI 2012, pp. 3121–3130 (2012)
9. Sakamoto, Y., Yoshikawa, T., Oe, T., Shizuki, B., Fukumoto, M., Tanaka, J.: Wrinklesurface: A wrinkleable soft-surfaced multi-touch interface. In: Proceedings of the 19th Workshop on Interactive Systems and Software, WISS 2011, pp. 7–12 (2011) (in Japanese)
10. Takeoka, Y., Miyaki, T., Rekimoto, J.: Z-touch: an infrastructure for 3d gesture interaction in the proximity of tabletop surfaces. In: ACM International Conference on Interactive Tabletops and Surfaces, ITS 2010, pp. 91–94 (2010)
11. Vlack, K., Mizota, T., Kawakami, N., Kamiyama, K., Kajimoto, H., Tachi, S.: Gelforce: a vision-based traction field computer interface. In: CHI 2005 Extended Abstracts on Human Factors in Computing Systems, CHI EA 2005, pp. 1154–1155 (2005)
12. 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)

Recognition of Multi-touch Drawn Sketches

Michael Schmidt and Gerhard Weber

Dresden University of Technology, Institute of Applied Science,
Human-Computer Interaction, Nöthnitzer Straße 46, 01062 Dresden
{Michael.Schmidt1, Gerhard.Weber}@tu-dresden.de

Abstract. We present concepts and possible realizations for the classification of multi-touch drawn sketches. A gesture classifier is modified and integrated into a sketching tool. The applied routines are highly scalable and provide the possibilities of domain independent sketching. Classification rates are feasible without exploiting the full potential of the scheme. We demonstrate that the classifier is capable of identifying common basic primitives and gestures as well as complex drawings. Users define sketches per templates in their individual style and link them to constructed primitives. A pilot evaluation is conducted and results regarding sketching techniques of users and classification rates are discussed.

Keywords: Sketch, recognition, classifier, survey, gestures, multi-touch.

1 Introduction and Motivation

A sketching software enhanced by methods for recognition allows for sketches to be interpreted, edited, searched and neatened [1]. Since its early years (e.g., ‘Sketchpad’ [2] introduced 1963), recognition of sketches reached manifold applications in different domains. Various sketching applications are for UML [3,4] and other diagrams [5,6,7], user interface design [8,9,10], mechanical schematical sketches [11], 3D curve modeling [12] and many more. Domain-independent approaches as in [13] exist, too. Additionally, gesture-based interfaces became ubiquitous in recent years. Mainly, their application is in direct manipulations for the scaling or re-orientation of graphical user interface elements. Nevertheless, new interaction techniques evolved in recent developments. Aiming to create more natural sketching interfaces, both techniques are often combined and switching of input modalities is required. We present a multi-touch sketching editor that extends sketch-based interaction techniques and circumvents the bounding to pen-based input. The gap between gestural interaction and sketching is alleviated by allowing to sketch with multiple fingers. Furthermore, domain-dependency is diminished by enabling the definition of primitives and complex multi-stroke symbols per examples.

2 Background and State of the Art

The subtasks of sketch recognition are pre-processing of strokes, the recognition of symbols or primitives¹ and their higher-level interpretation as sketches [15]. To recognize sketches, the segmentation of strokes² is crucial. When several objects are drawn, segmentation is complicated by the necessity to group strokes under unknown ordering and possibly incomplete specification (unconnected strokes) [17]. In such a scenario, segmentation involves the two stage preprocessing of selecting the correct stroke set and fragment or join selected strokes. One way to solve both tasks is to restrict the drawing styles of a user. In [18], these constraints and the training effort in consequence are stated to be reasons that sketch recognition has not yet entered into mainstream technology.

2.1 Conventional Approaches

Common approaches try to avoid drawing restrictions by sophisticated grouping [7,19] and segmentation [20,13,15,21,1] methods. Such **grouping of strokes** can be supported by additional constraints to input as time-outs or button presses to signal completion of a shape [14]. Predefined time-outs for collecting strokes to a scribble are used in [17]. In [7], polygonized strokes are connected if close (within a threshold) to another, while spatial and temporal closeness of strokes for grouping is exploited in [3]. Sezgin and Davis [19] apply grouping on a set of primitives gained by segmentation and recognition routines of [21] when analyzing complete scenes. They regard different drawing orders of primitives within symbols as they appeared on previous user observations.

If further **segmentation of strokes** is applied, curvature or speed information can be utilized. Calhoun et al. [15] derive segmentation points within strokes from relative (to stroke's input velocity) speed information and curvature. If speed falls under a certain threshold, a segmentation point is found. A similar method, based on thresholding velocity and directional changes within strokes, is applied by Sezgin et al. [21]. In [20], segmentation is done by turning points, detected by acceleration, speed and change of angle criteria.

These procedures are not necessarily strictly detached from the **recognition of primitives**. In [1], sophisticated segmentation of multi-stroke symbols into lines and arcs with dynamic programming techniques is applied. However, their approach needs pre-segmented templates of (unordered) primitives to find best matches between different segmentations of input and a template. Furthermore, as in most works, no merging of strokes is done and input is required to have fewer strokes than the template consists of. In [13], segmentation of a stroke is performed on points with highest curvature immediately after its drawing is finished. If an approximation of those segments to primitive shapes is not possible,

¹ In [14] primitives are seen as commonly used shapes or basic mathematical describable objects in a beautified form.

² Typically, sketches are produced by sequences of single-touch trajectories, commonly referenced as strokes. We adopt this terminology in this section for the ease of reading. In the context of gestures, terms defined in [16] avoid ambiguity.

this procedure is applied recursively to stroke segments. During post-processing, merging of primitives is applied if it leads to more meaningful primitives.

In recognizing primitives, mostly **rule-based approaches** are seen in literature, more or less formalized, ranging from ad hoc checks of geometric properties to decision trees and fuzzy rules. Apte et al. [17] compute convex hulls on their grouped and ordered strokes followed up by tests on geometric measurements (i.e., area/length ratios) with quasi decision trees. Fonseca et al. [22] improve on this approach and classify objects by manually generated fuzzy logic rules on percentiles of class specific sets of geometric features. Their use of global geometric properties in principle allows for multi-stroke input and dashed lines, but some composed primitives as arrows or crossed lines demand to be drawn partially in one stroke. Sezgin et al. [21] perform property checks to test their stroke segments first for lines and curves and afterwards for basic combinations as polylines, ovals, circles, rectangles and squares. In [19], besides parameterized (different slopes) versions of ovals and lines, checks for consecutive intersecting strokes are added. Hammond and Davis [3] pick up the work of [21] and include geometric property checks do detect arrows. Zeleznik et al. [5] extend the work of [23] and utilize the segmentation procedure of [15] and heuristic ad hoc rules on spatial and temporal properties to differ strokes in gestural commands, writing and primitive shapes. More rule-based methods are seen, for instance, in [20] (hierarchical fuzzy rules), [15] (semantic networks) and [13,18] (heuristic reasoning rules). In [24], Bayesian networks are applied and contextual information is used to support the interpretation of strokes.

2.2 Approaches Based on Gesture Recognition

The recognition of symbolic surface gestures is a very similar problem to that of sketch recognition. In many cases in literature, distinction of these two problems is neglected. Rubine's approach for gesture recognition [25], for instance, is used directly to detect gestures and primitive shapes in [8,9]. Furthermore, it is usually listed at comparing approaches for sketch recognition [26,13,18,11] though [18] criticizes the lack of drawing variability of such feature-based methods. Rubine's feature extraction routine is applied in [27] (which itself is integrated in [6]). Instead of the Bayesian classification scheme for Gaussian distributed features and common covariance matrices, support vector machines (SVM) are used for recognition. Further works apply the same techniques to sketches and gestures [4,5,13,22].

Approaches resembling gesture recognizers [28,9] avoid explicit segmentation and grouping and burden the users with restrictions to their drawing styles. Interestingly, Sezgin and Davis [19] state that drawing styles, though different for users, stay mainly consistent per user. If a recognizer supports specification of primitives by templates, it may allow individualization of sketching to users' style. Nonetheless, while several versions of a sketch may be specified by templates, drawing directions and orderings must be applied as in training data. Still, other advantages support the application of such approaches.

In contrast to most rule-based sketch recognizers [17,22,21,5,20,15,13] that are more intuitive and graspable to developers, gesture-based approaches often support training by templates [25,27,9] that can easily be done by users.

It is also indicated by [22,7] that trainable, statistical classifiers are superior to rule-based methods in terms of recognition performance. Wenyin et al. [7] compare rule-based approaches with SVM and artificial neural networks (ANN) to recognize basic primitives, stating SVM and ANN to be more suitable regarding performance in solving such problems. In [22], a comparison of trainable recognizers - K-Nearest Neighbors, Inductive Decision Tree, Naïve Bayes - favors the latter statistical approach³.

Often, sketch recognizers handle a restricted set of primitives, i.e., the objects regarded in sketching tools are composed of low order primitives. Some segment strokes into lines and arcs [1,4,15] for the purpose of beautification. Others add geometric primitives as ellipses, tri- and rectangles [17,7,20] to support object identification. Domain specific or complex shapes are composed out of these sets of primitives either by rules [7,8] or grammars [26,6,4,27], though other approaches as Hidden Markov Models [19] are seen.

Free and scalable specification of primitives and complex shapes by users in a single, template-based method may greatly enhance the functionality and versatility of a sketching software. We apply a new adaptable multi-touch gesture recognition technique to sketches.

3 This Work's Contribution

We integrated a trainable classifier for symbolic multi-touch gestures into a sketching application⁴. This allows users to specify their own sketches consisting of variable strokes and possibly multiple simultaneous contacts on the touch sensing surface. Beautified objects are constructed by the user and linked with a gestural command resembling a sketched input of these objects. The recognizer needs only few templates for training and one may suffice⁵ in many cases. Additionally, it is invariant against rotation, scaling and speed of input, but returns parameters regarding these attributes. Therefore, the beautified version of the sketch can be scaled, translated and rotated to best fit the input.

The sketching software not only adapts to users' drawing styles but also to different sketching domains, UML or UI Design, for instance. This neglects the necessity of defining domain objects in a special sketch or gesture definition language. No re-combination out of a fixed set of primitives is needed and the non-visual statistical recognition eliminates the need for helper strokes to support

³ Though it needed less training samples per class than the other methods, it still required an impractical amount of more than 40 samples to get above 90% recognition rate.

⁴ 'SkApp' is an Android sketching application and was developed during practical courses for students as well as an assignment work to improve its usability.

⁵ For comparison, SILK [8] uses 15-20 templates for each of its four primitive components [29].

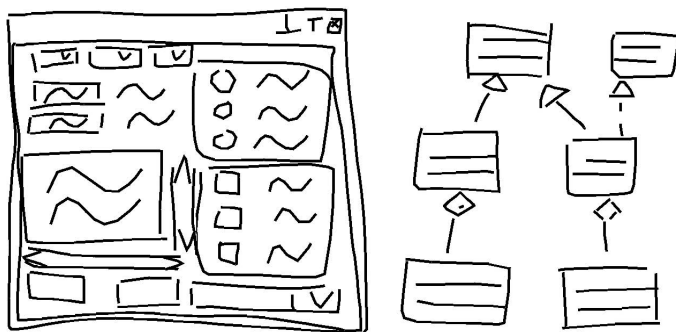


Fig. 1. A composition of uninterpreted (gestural) sketches. Arbitrary parts can be trained per templates and bound to an idealized shape. Each one may be specified with different numbers, orderings and concurrency of strokes and will be classified accordingly. If configured thus, the same idealized object is called to the canvas by sketches differing in those aspects.

merging. All sketches are defined by a collection of strokes, their shapes and relations to another. Sketching variability can be produced by specifying several different multi-touch, multi-stroke gestures⁶ per primitive. Figure 1 shows sketches of the two domains - GUI design and UML - we chose for the proof of concept for our domain independent multi-touch sketch recognition. The sketches were drawn by fingers. The two wavy lines within the rectangle, for instance, can be drawn simultaneously or sequentially (or in any temporary offset) and sketch templates may be specified in one way or another or both.

Additionally, users do not have to stick to sketching, they can also define short cut gestures, not necessarily similar to the sketch, to call complex shapes on the canvas. Different symbolic commands (sketches/gestures) can be specified for one shape to allow for variations in input. In the left sketch of Fig. 1, abstractions for the minimize and maximize buttons of the sketched window were specified. In the UML class diagram (right side of Fig. 1), the colored version of the rhombus for sketching a composition association instead of an aggregation is indicated by an additional single tab into the arrowhead. Such leeways might support the ease of sketching. When sketching by fingers, fluent transitions to direct manipulations are possible without changing the input device. New options for separating input modes, i.e., direct manipulation, sketching and command gestures, arise by utilizing different numbers of contacts for different modes. Also, the enhanced input variety allows for distinction of gestures and sketches by disjunct sets of templates. Either way, other techniques, such as the often applied modus buttons or time-outs, are still applicable.

We demonstrate that our classifier is capable of recognizing common basic primitives and gestures while allowing multi-touch input, too. The disadvantage of burden the user with explicit segmentation by strokes is compensated by its

⁶ In [14], it is stated that multi-stroke primitives are more prevalent in sketching than single-stroke ones. Multi-touch input may be beneficial, too.

flexibility for manifold input and its capacity to recognize a large set of thus composed complex drawings.

4 Realization of Sketch Recognition

We applied the on-line feature based gesture recognition routine of [16] in its unparameterized variant (see pseudocode in [30]). It implies a standard normal distribution of all features of a gesture. Therefore, the classification process mostly uses basic distance calculations. Furthermore, we do not exploit temporary features to allow for different stroke orderings in an input without explicitly defining additional templates.

While reviewing current literature on sketching, we came across a set of common primitives recognized by sketching tools, i.e., lines, arcs, ellipses, circles, triangles, rectangles, diamonds, arrows and crossing lines. Those types can be specified by templates to our recognizer and it is capable of distinguishing shapes of a much larger set and many variations. Other elements supported by some tools are helices, spirals, squiggly or wavy lines, scribble and lasso gestures. It is in the nature of our template-based approach that such primitives can not be handled in a straightforward way. Each type represents a group of elements that arbitrarily vary in their number of loops, turns or edges. Multiple templates might be specified regarding such modifications to cover the expected cases. Alternatively, this types of gestures have to be identified separately. This applies to another type of elements that are used in sketching tools. Normally, polylines, polygons and Bezier curves can be produced by re-sampling or smoothing, possibly supported by special detecting of segmentation points ([21]). Our pattern matching method is not suitable for detecting thus generalized primitives without specifying concrete versions, for instance, a rectangle. In our sketching editor ‘SkApp’, these primitives are constructed within special modes. Additionally, the nearest neighbor template matching differs between gestures in various modifications of their size in one dimension. We apply non-uniform scaling in cases an input is compared with a template representing shapes that differ only in the length of one dimension (sketches defining scrollbars, for instance).

5 Evaluation Routine

We conducted a user evaluation within the topics of designing UML diagrams and UI interfaces. Our intention was to elaborate on the question of performance and usability of our sketch input method as well as the users’ behavior in exploiting their broader range of possibilities. Figure 2 shows the sketching application ‘SkApp’ and interpreted sketches as those demonstrated in Fig. 1.

The application contains in the upper part a row of buttons, each representing a single primitive. By a long tap on a button, the primitive together with the drawing of a gesture for calling it onto the canvas is shown. On the right side, options for coloring and selecting stroke types are available. We specified the 19 primitives of Fig. 2 for user interface elements (button, checkbox, close,

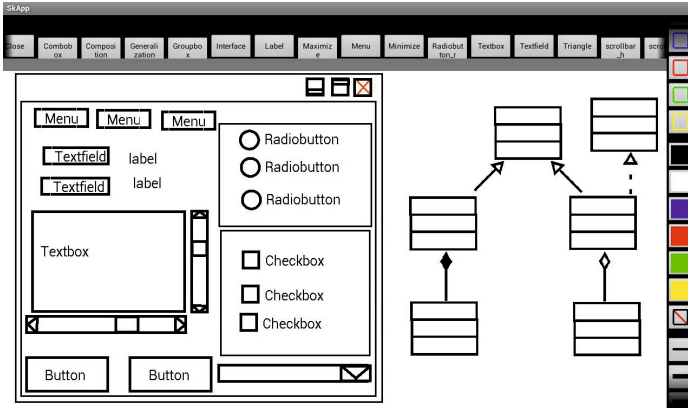


Fig. 2. Interpreted sketches of GUI and UML. Elements were entered separately by gestures within a sketching mode. Different gestures were specified for each graphical element in advance.

combobox, groupbox, label, maximize, menu, minimize, radiobutton, horizontal scrollbar, vertical scrollbar, textbox, textfield) and class diagrams (aggregation, class, composition, generalization, interface) together with associated gestures. The primitives were constructed within 'SkApp' and parameters determining how they should respond to input variations regarding scale and rotation were configured. UI elements were defined not to rotate at all and most elements do not scale (button, checkbox, close, label, maximize, menu, minimize, radiobutton) while some scale in one dimension only (combobox, horizontal scrollbar, vertical scrollbar, textfield) and others in both (groupbox, textbox). All relationships of class diagrams are allowed to scale and rotate while a class itself stays in one size and one orientation only.

At specifying gestures, we assumed a general top-down and left-right drawing direction. Rectangle parts, for instance, can be drawn by a single stroke beginning in the upper left corner and drawing in the two possible directions or a multi-stroke variation with two strokes drawn top-down and two left to right. As we excluded temporal features in our classification routine, multi-stroke gestures can be performed in any possible temporal arrangement. Visually, all assigned gestures correspond to the variations indicated in Fig. 1. Overall, two sets of 28 gestures for the GUI elements and 17 gestures for UML were predefined. Each gesture represents one variant of an element.

Note that our classification is invariant to rotation and scale even if sketches are configured not to scale or rotate with input. For few elements (scrollbars, textfield, combobox), non-uniform scaling is applied in advance to the general template matching routine. Primitive dependent restrictions to this invariances are expected to improve recognition rates significantly. Likewise, an improvement can be achieved by specifying more than one template per variation and specifications done by users only.

5.1 Testbed

We asked 8 participants - 4 male, 4 female, all educated in applied computer science, familiar with UML and GUI widgets, and right handed - to draw the sketches of Fig. 2. The evaluation was done on an Android Tablet of type Motorola Xoom. To demonstrate the concepts, all participants got a cheat sheet showing the primitives and one assigned gesture for each primitive. The pictures did not convey the drawing directions of our templates and the participants were encouraged to apply their own drawing style. It was mentioned that it is allowed to utilize input by multiple fingers. If participants attempted to draw a primitive in a way not covered by our templates, they were instructed to specify an own version of the gesture. In case of misinterpretations or new definitions, input was repeated and tests ended on complete sketches. A thinking aloud evaluation was applied to know in advance which type of primitive a participant intended to draw. We collected data on drawing styles, misclassification rates and added gesture types.

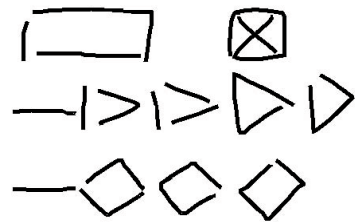
5.2 Results

Figure 3 shows results of our evaluation regarding misinterpretations and definitions of additional templates.

Re-definitions were necessary to draw ‘Close’ (one user) and arrowheads by two strokes (two user for the triangle head type and two for the rhombus). Two users (Chinese) had to redefine all elements containing rectangles as they drew them by three strokes (see right side of Fig. 3). Few other re-definitions were in respect to the directions of strokes, for instance, for the x within the ‘Close’ widget (1 time) and to the starting point of single-stroke arrowheads (original

mis-interpretations at 208 + 88 entered elements	Button	Checkbox	Groupbox	Radiobutton	Textfield	Aggregation	Generalization	by participants
Button			8					5
Checkbox				8				4
Combobox					1			1
Groupbox	4							3
Radiobutton		2						2
Scrollbar	1							1
Aggregation							2	1
Generalization						5		2

(a) misinterpretations



(b) redefinitions

Fig. 3. Misinterpretations (table) of input (row) and the number of participants for which they were observed. On the right side are representatives of the templates specified by users.

templates regarded one starting point only). For three users, the latter aspect seemed to depend on the arrow's direction. One user drew the strokes of the scrollbar in a different direction and the shaft of arrows away from their head. Though two users remarked that they would prefer input by pen to avoid occlusions by fingers, multi-touch input was applied rather consequently. One user did not utilize multi-touch at all. Two users did multi-touch input for scrollbars only, four for the inner lines of the textbox and the class, three of them additionally for scrollbars. One user only drew the outer boxes of the textbox and the combobox by multi-touch.

Overall, our classifier achieved a recognition rate of 88.5% for the GUI elements and 92% for UML.⁷ Ignoring confusions of buttons and groupboxes - both were specified as rectangles differing in their relation of height and with only - the rate increases to 94.2% for the first set. Note that participants had to repeat an input until finalization of their sketch. Additionally taking out this fact would result in slightly higher rates.

5.3 Discussion

Though different users applied different drawing styles, the style of a user was very consistent. This confirms observations of Sezgin and Davis [19]. Two Chinese participants indicated cultural differences in drawing styles as they were accustomed to draw all rectangles by three strokes. Results in [10] showed that users vary in their ideas of how to draw UI widgets. We conclude that a sketch recognition should be adaptable to visual representations as well as users' drawing styles to produce them. This can be realized by a sketching software that allows for supplementing individual training samples to predefined templates. The most critical aspect is that starting points for drawings apparently changed depending on the rotation of an object. In our cases, this happened for sketches of arrows.

When comparing our classification results to others reported in publications, few works are suitable. Some authors do not provide tests at all [6,11,9]. Others perform informal tests, guess recognition rates [15,20], or do not provide sufficient information [4]. A comparison of three approaches can be found in [7], but tests are for three basic primitives (triangle, rectangle, ellipse) only. Yu and Cai [13] report 98% correct segmentation and a recognition rate of 70% over an imprecisely described set of shapes when accepting interpretations as soon as the main and key structures are recognized properly.

Sezgin et al. [21] achieved 96% correct segmentation in lines and curves on a set of 10 distinct shapes, but no recognition rates of their rule-based approach are reported. However, a HMM approach achieved 96.5% on a set of 10 complex shapes of different domains by training six styles, each with 10 examples, per shape [19]. Apte et al. [17] report 98% correct classifications of six types of shapes by their heuristic rules. The similar method of [18] correctly classified

⁷ Though with different numbers of test cases for each element.

98.5% of 8 primitive and complex shapes. In [22], 95.8% correct classification for 12 primitives are achieved by fuzzy logic.

A test setting comparable to ours is used in [29] where GUI mock-ups are sketched. Correct classification of 69% of the input is reported. The authors use Rubine's classifier trained with 15-20 templates per basic primitive and rules to identify composed elements. In [31], the rules of the recognition routine are modified to statistical disambiguation. It is indicated that results similar to ours are possible, albeit large training sets are required.

6 Critic and Future Work

The method presented in this work achieves feasible classification rates and can easily be developed into flexible and scalable tools. However, there are drawbacks. The approach presented needs explicit segmentation of input by the user. Though the style of drawing does not vary much in most cases, for some elements it can depend on the rotation of their input. If in this case elements are too complex, many variations within the template set are required. To overcome this, automatic segmentation as in [21] or [1] might be a useful instrument. Additionally, each comparison of a token - for shape and structure - can be done in both possible directions, choosing the one with minimum distance. In this way, only a few templates might again be sufficient and training of sketches stays comfortable. More issues are of concern when embedding our concepts into a usable tool. In our prototype application, sketches were entered within a mode kept by touching the surface with two fingers of the left hand. In this way, the user is responsible for the grouping of input, too. One possible improvement is to let users select elements per lasso gesture for an interpretation. This would allow to sketch more than one element or a complete design in a row. Interpretation can be done on explicit command. Other options are to apply grouping strategies or expensive searches. Furthermore, the sketching software should prevent users from defining too similar gestures for different objects. Finally, the selection of one of possible multiple interpretations or the complete rejection of an input for classification are to be included, though the latter is only a matter of setting a threshold.

Acknowledgments. Students of several internships were involved in the implementation of the sketching software 'SkApp' and its increasing functionalities. In alphabetical order, we want to thank: Björn Bussewitz, David Götze, Thilo Gürtler, Susanne Haase, Sascha Huth, Frederik Schulz, Christoph Senkel. We thank Dana Hank for optimizing the user interface and including bi-manual control. Last but not least, the authors kindly thank the participants of the user study for their support and helpful feedback.

References

1. Hse, H., Shilman, M., Newton, A.R.: Robust sketched symbol fragmentation using templates. In: Proceedings of the 9th International Conference on Intelligent User Interfaces, IUI 2004, pp. 156-160. ACM, New York (2004)

2. Sutherland, I.E.: Sketch pad a man-machine graphical communication system. In: Proceedings of the SHARE Design Automation Workshop, DAC 1964, pp. 6.329–6.346. ACM, New York (1964)
3. Hammond, T., Davis, R.: Tahuti: A geometrical sketch recognition system for uml class diagrams. In: Papers from the 2002 AAAI Spring Symposium on Sketch Understanding, Stanford, California, USA, pp. 59–68. AAAI Press (2002)
4. Costagliola, G., Deufemia, V., Risi, M.: A trainable system for recognizing diagrammatic sketch languages. In: 2005 IEEE Symposium on Visual Languages and Human-Centric Computing, pp. 281–283 (September 2005)
5. Zeleznik, R.C., Bragdon, A., Liu, C.C., Forsberg, A.: Lineogrammer: creating diagrams by drawing. In: Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology, UIST 2008, pp. 161–170. ACM, New York (2008)
6. Jansen, A., Marriott, K., Meyer, B.: Cider: A component-based toolkit for creating smart diagram environments. In: Blackwell, A.F., Marriott, K., Shimojima, A. (eds.) Diagrams 2004. LNCS (LNAI), vol. 2980, pp. 415–419. Springer, Heidelberg (2004)
7. Wenyin, L., Qian, W., Xiao, R., Jin, X.: Smart sketchpad - an on-line graphics recognition system. In: Proceedings of the Sixth International Conference on Document Analysis and Recognition, ICDAR 2001, p. 1050. IEEE Computer Society, Washington, DC (2001)
8. Landay, J.A., Myers, B.A.: Interactive sketching for the early stages of user interface design. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 1995, pp. 43–50. ACM Press/Addison-Wesley Publishing Co., New York, USA (1995)
9. Coyette, A., Schimke, S., Vanderdonckt, J., Vielhauer, C.: Trainable sketch recognizer for graphical user interface design. In: Baranauskas, C., Abascal, J., Barbosa, S.D.J. (eds.) INTERACT 2007. LNCS, vol. 4662, pp. 124–135. Springer, Heidelberg (2007)
10. Caetano, A., Goulart, N., Fonseca, M., Jorge, J.: Javasketchit: Issues in sketching the look of user interfaces. In: AAAI Spring Symposium on Sketch Understanding, pp. 9–14. AAAI Press, Menlo Park (2002)
11. Kurtoglu, T., Stahovich, T.F.: Interpreting schematic sketches using physical reasoning. In: Proceedings of the AAAI Spring Symposium, Stanford, CA, pp. 78–85. AAAI Press (2002)
12. Bae, S.H., Balakrishnan, R., Singh, K.: Ilovesketch: as-natural-as-possible sketching system for creating 3d curve models. In: Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology, UIST 2008, pp. 151–160. ACM, New York (2008)
13. Yu, B., Cai, S.: A domain-independent system for sketch recognition. In: Proceedings of the 1st International Conference on Computer Graphics and Interactive Techniques in Australasia and South East Asia. GRAPHITE 2003, pp. 141–146. ACM, New York (2003)
14. Hammond, T., Paulson, B.: Recognizing sketched multistroke primitives. ACM Trans. Interact. Intell. Syst. 1(1), 4:1–4:34 (2011)
15. Calhoun, C., Stahovich, T.F., Kurtoglu, T., Kara, L.B.: Recognizing multi-stroke symbols. In: Proceedings of the AAAI Spring Symposium - Sketch Understanding, pp. 15–23. AAAI Press (2002)
16. Schmidt, M., Weber, G.: Template based classification of multi-touch gestures. Pattern Recognition (2013), doi:10.1016/j.patcog.2013.02.001, <http://www.sciencedirect.com/science/article/pii/S0031320313000770>

17. Apte, A., Vo, V., Kimura, T.D.: Recognizing multistroke geometric shapes: an experimental evaluation. In: *UIST 1993: Proceedings of the 6th Annual ACM Symposium on User Interface Software and Technology*, pp. 121–128. ACM, New York (1993)
18. Paulson, B., Hammond, T.: Paleosketch: accurate primitive sketch recognition and beautification. In: *Proceedings of the 13th International Conference on Intelligent User Interfaces, IUI 2008*, pp. 1–10. ACM, New York (2008)
19. Sezgin, T.M., Davis, R.: Hmm-based efficient sketch recognition. In: *IUI 2005: Proceedings of the 10th International Conference on Intelligent User Interfaces*, pp. 281–283. ACM, New York (2005)
20. Chen, C.L.P., Xie, S.: Freehand drawing system using a fuzzy logic concept. *Computer-Aided Design* 28(2), 77–89 (1996)
21. Sezgin, T.M., Stahovich, T., Davis, R.: Sketch based interfaces: Early processing for sketch understanding. In: *Workshop on Perceptive User Interfaces, Orlando FL* (2001)
22. Fonseca, M.J., Pimentel, C., Jorge, J.A.: Cali: An online scribble recognizer for calligraphic interfaces. In: *Sketch Understanding, Papers from the 2002 AAAI Spring Symposium*, pp. 51–58 (2002)
23. Igarashi, T., Matsuoka, S., Kawachiya, S., Tanaka, H.: Interactive beautification: a technique for rapid geometric design. In: *Proceedings of the 10th Annual ACM Symposium on User Interface Software and Technology, UIST 1997*, pp. 105–114. ACM, New York (1997)
24. Alvarado, C., Oltmans, M., Davis, R., Davis, A.: A framework for multi-domain sketch recognition. In: *AAAI Spring Symposium on Sketch Understanding*, pp. 1–8. AAAI Press (2002)
25. Rubine, D.: Specifying gestures by example. *SIGGRAPH Comput. Graph.* 25(4), 329–337 (1991)
26. Hammond, T., Davis, R.: Ladder, a sketching language for user interface developers. In: *ACM SIGGRAPH 2007 Courses*. ACM, New York (2007)
27. Bickerstaffe, A., Lane, A., Meyer, B., Marriott, K.: Graphics recognition. In: *Recent Advances and New Opportunities*, pp. 145–156. Springer, Heidelberg (2008)
28. Rubine, D.: *The Automatic Recognition of Gestures*. PhD thesis, Carnegie Mellon University (1991)
29. Landay, J.A., Myers, B.A.: Sketching interfaces: Toward more human interface design. *Computer* 34(3), 56–64 (2001)
30. Schmidt, M., Fibich, A., Weber, G.: MTIS: A Multi-Touch Text Input System. In: *Streitz, N., Stephanidis, C. (eds.) DAPI/HCI 2013. LNCS, vol. 8028*, pp. 62–71. Springer, Heidelberg (2013)
31. Shilman, M., Pasula, H., Newton, S.R.R.: Statistical visual language models for ink parsing. In: *Proceedings of the AAAI Spring Symposium on Sketch Understanding*, pp. 126–132 (2002)

A Web Browsing Method on Handheld Touch Screen Devices for Preventing from Tapping Unintended Links

Yu Shibuya, Hikaru Kawakatsu, and Kazuyoshi Murata

Kyoto Institute of Technology, Kyoto, Japan
{shibuya, kmurata}@kit.ac.jp

Abstract. In recent years, it is common thing to browse Web pages with mobile devices, such as smart phones. However, users sometimes tap the wrong link when they scroll or zoom web pages because of the relatively small display area of mobile device and sensitivity of touch screen. In such case, it is necessary to stop of loading the page or back to the previous page after the page changed. It seems that above unintended operation might increase the total browsing time and user's frustration. In this study, we aimed to prevent users from tapping unintended links for effectively web browsing with touch-screen mobile devices. The proposed method has two kind of operation mode. They are a tapping mode and non-tapping mode. With the tapping mode, users can tap the link and change the mode only. On the other hand, with the non-tapping mode, users can do swipe, pinch, and mode change operation but they cannot tap any links. Furthermore, mode change operation, we adopt the Bezel Swipe operation, is intuitive and efficient.

The results of the experimental evaluation showed that the rate of tapping the unintended links with the proposed method was lower than that with conventional method. However, the task completion time with proposed method is longer than that with conventional method.

Keywords: Mobile interaction, web browsing, unintentional tap, Bezel Swipe.

1 Introduction

In recent years, it is common thing to browse web pages with touch screen mobile devices, such as smart phones. However, users sometimes tap the wrong link when they scroll or zoom in or out web pages because of the small screen area of mobile device and sensitivity of touch screen. In such case, it is necessary to stop of loading the page or back to the previous page after the page changed. These unintended operations might increase the total browsing time and user's frustration.

Matero and Colley analyzed accidental touches on capacitive touch screen based mobile telephones in a user test [2]. In the study, patterns that are characteristic of unintentional touches were identified and layout guide lines to reduce the amount of them were presented.

In this study, we aimed to prevent users from tapping unintended links for efficiently web browsing with touch screen mobile devices.

2 Proposed Method

The proposed method has two kinds of operation mode. They are a tapping mode and a control mode. With the tapping mode, users can tap the link and change the mode only. On the other hand, with the control mode, users can do scroll, zoom in or out, and mode change operation but they cannot tap any links. Furthermore, mode change operation, we adopt the Bezel Swipe operation [1], is intuitive and efficient. As shown in Fig. 1, the Bezel Swipe is done by sliding of the thumb or finger, starting from a bezel zone to a screen zone of the device.

When users want to change the mode from tapping to control, they can do that with just swipe in their finger from the outside of the screen. In order to change the mode from control to tapping, users can do that with just tapping on the screen.

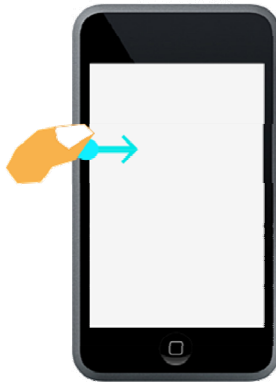
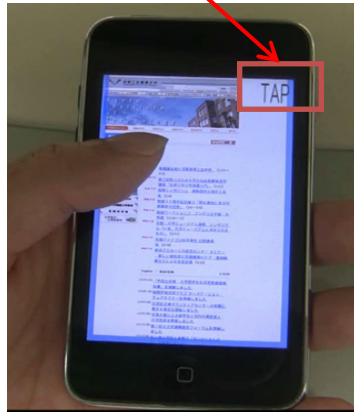


Fig. 1. Bezel Swipe

In order to distinguish the tapping mode from control mode, there is a visual feedback on the top right of screen. As shown in Fig. 2, a word “TAP” is displayed at there in the tapping mode or a word “CTRL” is displayed in the control mode.

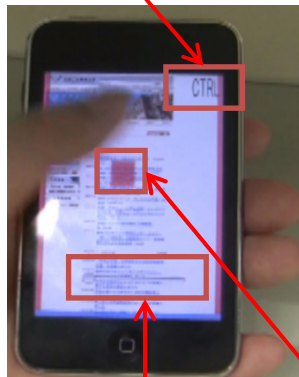
In control mode, two kinds of operation method are proposed for zooming in this paper. One is called as “double-tap method” and another is called as “slider method”. In double-tap method, users can zoom in or out by double tapping on the screen. The number of zooming level is two, so a double tapping toggles the zooming level between zoom-in status and zoom-out status. In zoom-in status, the width of tapped area is adjusted to the screen width as the common web browser’s zoom-in manner.

A visual feedback of tap mode



(a) tapping mode

A visual feedback of control mode



A finger take-off point

A slider for zooming

(b) control mode (slider method)

Fig. 2. Two kinds of mode in proposed method

While the zooming level is discrete in double-tap method, users can set the zooming level continuously with the slider method. In the slider method, there is a slider at the bottom of the screen and users can slide the tab of it to set the zooming level as they want.

Furthermore, in slider method, in order to make the operation efficient, there are two kinds of operation manner depends on the bezel swipe direction to change the mode. When users swipe in their finger from the bottom of the screen, the mode is immediately set the control mode and they can scroll the page with swipe action without finger taking off. On the other hand, when users swipe in their finger from either side of the screen, the mode is set the control mode and the point where they take off their finger is the center for zooming in or out operation. It is expected that they will take off their finger where they want to see in detail. Then they can set the zooming level with the slider as they want.

3 Experimental Evaluation

In the experiment, three kinds of web browsing method on handheld touch screen devices are compared. They are a conventional method, the double-tap method, and the slider method. The conventional method is a commonly used modeless method on handheld touch screen devices. Users can scroll pages with swipe, zoom in or out with double tapping, and change a page with tapping a link.

Our proposed method is compared with this conventional method experimentally. Both proposed and conventional methods are implemented on the Apple's iPod touch. A repeated measurements within-subject design is used for the experiment.

The purpose of the experiment is to examine following hypotheses:

- H1: Both proposed method, they are the double-tap method and the slider method, decrease the rate of unintended tapping compared with conventional method.
- H2: Both proposed method slightly but not significantly increase the operation time compared with conventional method because there is additional operation, which is mode change operation.

3.1 Web Pages for Task

For the experiment, we prepared top pages of the Yahoo! JAPAN web site for PC, the Wikipedia Japan web site, and our university's web site. From each web page, ten different web pages are made with making different link's string to be red. There were total thirty pages were prepared for the experiment. An example web page the experiment is shown in Fig. 3.



Fig. 3. An example web page of the experiment

3.2 Procedure

Ten participants were recruited from our university. In the experiment, each participant was asked to hold the mobile device with their one hand and to operate it with the same hand thumb.

Each participant was asked to do following procedure to complete the task for each method:

1. Do practice until enough for operation.
2. Touch the start button on the screen. Then a web page is displayed.
3. Touch the link with red string. Then a new web page is displayed.
4. After all designated web page is touched, and then the “task completed” message is displayed.

Error rate was calculated with dividing the number of web page which had unintended user’s touch by the number of total web page for the experiment. Task completion time, the elapsed time from pressing the start button to displaying the “task completed” message, was also measured. Furthermore, subjective evaluation was done with a questionnaire after each task.

3.3 Results and Discussion

The results of the experimental evaluation showed that there was a main effect of the method on the error rate ($F(1,553)=11.824, p<0.05$). The error rate of tapping the unintended links with the double-tap method was 3.33[%] and that with the slider

method was 1.67[%]. There was no significant difference between them but both error rate was significantly lower than that of conventional method (9.17[%], $p < 0.01$). This result supports our hypothesis H1.

On the other hand, there was a main effect of the method on the task completion time. The time with the double-tap method was 210.89[sec] and that with the slider method was 227.83[sec]. There was also no significant difference between them but both task completion time was significantly longer than that of conventional method (157.12[sec], $p < 0.05$). This result does not support our hypothesis H2.

Results of the subjective evaluation showed that participants viewed the proposed method very positively. Compared with the conventional method, the proposed method got the same or nearly the same score about the easiness to learn the usage and the easiness of operation.

4 Conclusion

In this paper, a web browsing method on handheld touch screen devices for preventing from tapping unintended links is proposed and evaluated experimentally. From the experiment, it is found that the proposed method is decrease the rate of unintended tapping compared with the conventional method. However, it is not so efficient for operation on mobile devices. From the subjective evaluation, it is also found that the proposed method is viewed very positively.

References

1. Roth, V., Turner, T.: Bezel swipe: conflict-free scrolling and multiple selection on mobile touch screen devices. In: CHI 2009 Proceedings of the 27th International Conference on Human Factors in Computing Systems, pp. 1523–1526 (2009)
2. Matero, J., Colley, A.: Identifying unintentional touches on handheld touch screen devices. In: Proceedings of the Designing Interactive Systems Conference, pp. 506–509 (2012)

Real Time Mono-vision Based Customizable Virtual Keyboard Using Finger Tip Speed Analysis

Sumit Srivastava and Ramesh Chandra Tripathi

SILP Lab, Indian Institute of Information Technology, Allahabad
srv.sumit@gmail.com, rctripathi@iiita.ac.in

Abstract. User interfaces are a growing field of research around the world specifically for PDA's, mobile phones, tablets and other such gadgets. One of the many challenges involved are their adaptability, size, cost and ease of use. This paper presents a novel mono-vision based touch and type method on customizable keyboard drawn, printed or projected on a surface. The idea is to let the user decide the size, orientation, language as well as the position of the keys, a fully user customized keyboard. Proposed system also takes care of keyboard on uneven surfaces. Accurate results are found by the implementation of the proposed real time mono-vision based customizable virtual keyboard system. This paper uses a phenomenal idea that the finger tip intended to type must be moving fastest relative to other fingers until it does a hit on a surface.

Keywords: Virtual Keyboard, Image Processing, Single camera, mono vision, Edge Detection, Quadrilateral extraction, Character Recognition, Hand Segmentation, Fingertip extraction, Customizable keyboard.

1 Introduction

This paper introduces a method to untangle several issues related to input through keyboard such as portability, cost, shape and size of the keyboard, orientation, position of keys, multilingual support. Multilingual support is one of the most important aspects due to large number of languages around.

Significant amount of research is being done in the domain of user interfaces using computer vision but only a few exists which talk about single camera based techniques. This certainly brings down the resources used and thus more cost effective and portable device.

Habib et al. [1] proposed a mono-vision fuzzy rules based technique which has some pre-defined rules to identify a particular gesture of fingers as action to press a particular key. The method involves a SDIO camera being placed at table top and records the movement the finger tips and knuckles. It does so for a particular number of frames and then decides which keystroke that particular hand gesture may be for. It pre-defines some 32 fuzzy rules for those different finger tip and finger knuckles position. This method is quite good but is not flexible to different keyboard patterns. The rules are fixed and cannot be used for all language keyboards due to varying number of alphabets and keyboard layouts.

Murase et al. [2] has proposed a method involving the camera being placed on the table top such that its view tends to be in parallel to the table surface. It looks out for fingertips and identifies a finger tip movement as a keystroke if it descends down below a particular level in the frame. It calculates the depth of finger using real-AdaBoost machine learning algorithm that adopts HOG features of user's hand image. This work has a limitation that the keyboard has to be pre-defined with exact depth knowledge of each key. Again it faces the problem that it is not so robust with any keyboard and also not so user friendly as keyboard depth will have to be defined by the user.

Adajani et al. [3] proposed a technique that uses the idea that a finger tip and its shadow shall coincide when the key touch event happens. This technique falls short on light source point of view. Likewise, if the light source is not exactly in front of the hand, this shall fail. Also a problem arises when two fingers are very close to each other and are very near to the surface, resulting in a situation such that shadow of one is absorbed by the other.

Conclusively most of the methods face the problem of pre-defined keyboards, i.e., information about keys position, orientation, language, size, distance from camera, shape of the keyboard should be known. Also some are completely dependent on light source position.

We thus try to solve all the above problems with urge to make it more robust and user friendly.

2 Proposed Methodology

The implementation flow graph is shown in Fig.1. The first frame captured is used to localize the keys. In the next phase these identified possible key locations are passed

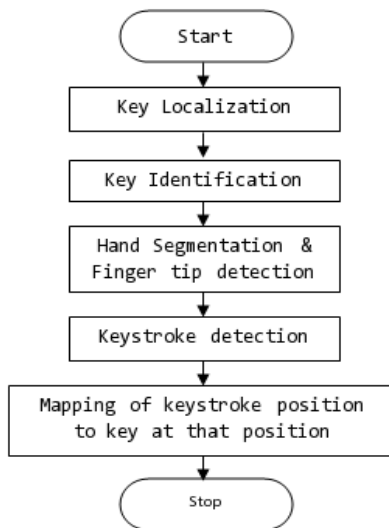
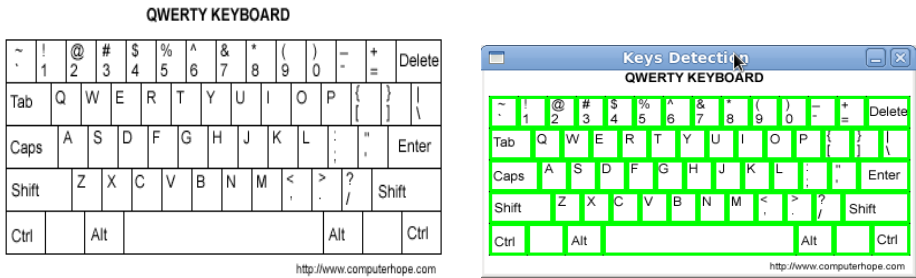


Fig. 1. Flow chart of the proposed methodology

through an optical character recognition procedure to identify the keys. Next, the user hand is segmented using colour based hand segmentation and subsequently the fingertips are found. Then, we detect the key touch events by tracking the fastest moving finger tip in subsequent frames. In the end, the touch points are mapped to the key locations identified in first phase. Following is the detailed discussion of all the phases.

2.1 Key Localization

This phase deals with localization of keys, i.e., identifying position of the keys. For this, the first frame from the camera is obtained having the keyboard drawn/ printed/ projected on a surface. Image contrast enhancement is done using Gray-Level-Grouping [4]. All the contours are detected from the image after running canny edge detector on it and also by using several pre-defined threshold levels and removing end points that have no connectivity [5]. All the contours having number of sides as four, an area in a specific range and being convex are filtered out from the large pool of contours obtained. Angle is measured between the sides of these filtered polygons. Those with cosines greater than 0.6, are rejected. The allowed range of cosines is taken large because certain genuine quadrilaterals also have irregular shape. Fig. 2 shows some results of the first phase obtained in several conditions.

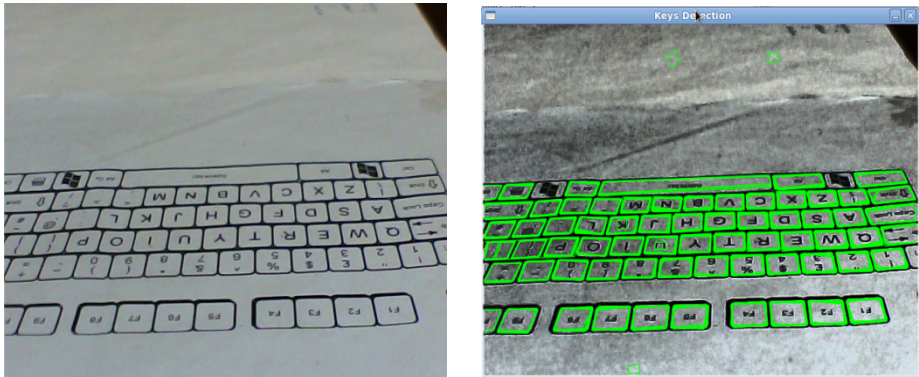


[a]



[b]

Fig. 2. [a] Computer generated keyboard and its result, [b] Keys drawn on a newspaper and its result, [c] Keyboard printed on a sheet of paper and its result



[c]

Fig. 2. (Continued)

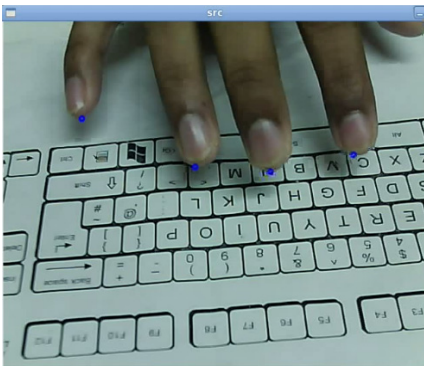
2.2 Key Identification

Key identification process involves implementation of Optical Character Recognition. There are several good OCR algorithms around. Here we have used the method proposed by Gupta et. al [6] for hand written character recognition using neural network. The method classifies characters by two approaches, holistic and segmentation based. We have used holistic method as most keys obtained are individual characters and thus further segmentation is not required. Also those keys having more than one character are limited in number and only slight variation is present. Further, Fourier descriptors are used as features and the classification is done using several classifiers, MLP, RBF and SVM. As of the proposed approach SVM provides best results, so we have used it directly.

2.3 Hand Extraction

There are several moving object extraction methods such as background subtraction [7 - 8], temporal difference, optical flow analysis [9], Gaussian model, HOG, force field method [10], etc. Of these only temporal filtering is one which can be readily used in real time systems. Modifications of these approaches are also available which can provide real time results which are quite good to work upon [11 - 13]. But here we have used the simple HSV colour space based segmentation for skin [14] which certainly puts on some constraint on the environment it can be used in, such as shadow could be a problem in some cases where light source is not directly in front of the hand. Such issues can be handled by using better segmentation techniques involving removal of shadow and ghost effect.

This segmentation is then followed by exclusion of unwanted areas segmented due to similar colour. This is done by excluding segments smaller than a particular size. Also some post-processing is applied as closing operation to enhance the segmented hand as seen in figure 3[a] and 3[b].



[a]



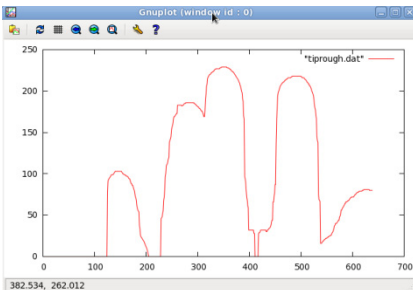
[b]

Fig. 3. [a] Source hand image. [b] Segmented hand from the source image.

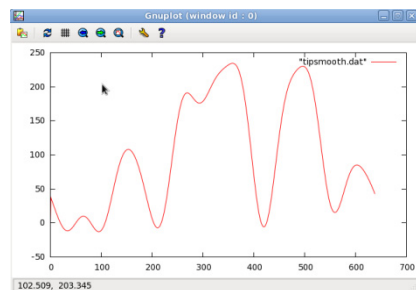
2.4 Finger Tip Extraction

Hand segmentation provides a rough estimation of finger tips. To extract them, a process inspired from the star skeletonization method [16] is used. Here, a horizontal scanning of the segmented image is done to obtain the highest “lighted” pixel on the y-axis. This provides an outline of the segmented hand.

Further, a DFT is applied on this obtained curve followed by a low pass filter and an IDFT to obtain a smooth curve highlighting the finger tips in form of local maxima for each tip. These maxima are then synchronized with the proper tip it corresponds to. This phase gives a decent result with all tips being found.



[a]



[b]

Fig. 4. [a] Rough finger outline, [b] Smooth curve obtained after applying DFT followed by Low pass filter and IDFT with local maxima representing tips.

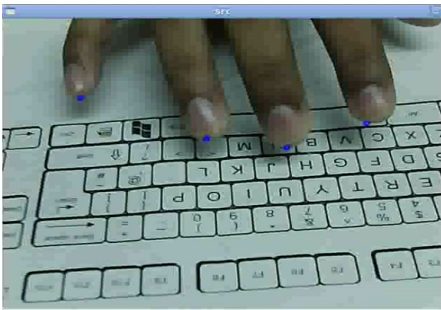
2.5 Keystroke Detection

This phase follows up to the previous one with tips being found in every frame and maintaining location of each tip in each frame with several other data such as direction of the current motion, last point of change of direction, base frame from which movement in a particular direction began.

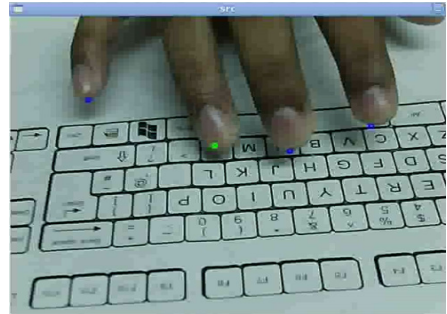
All of this data is used to keep track of the speed of each tip in each frame. This is calculated by the following formula,

$$S_i = \frac{(CP_i - PBF_i)}{N}, \quad (1)$$

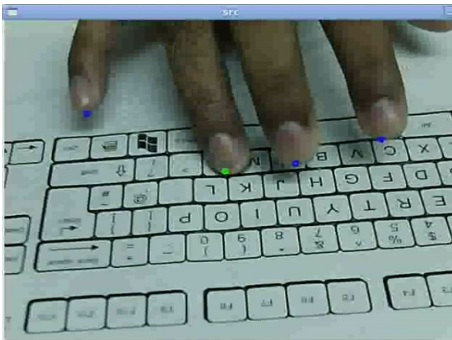
S_i represents speed of the tip in current frame, CP_i represents position of the tip in current frame, PBF_i is for position of the tip in base frame, N represents number of frames between current frame and base frame for the particular finger tip.



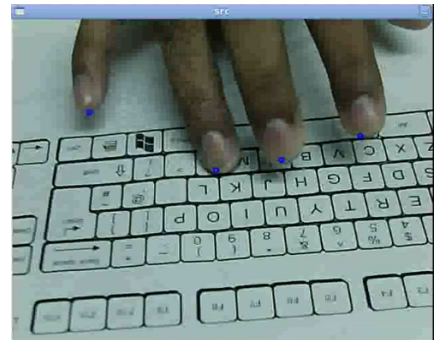
[a]



[b]

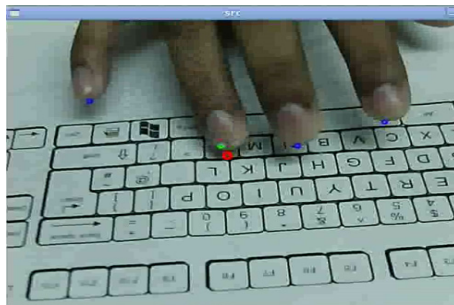


[c]



[d]

Fig. 5. [a]-[e]. Subsequent frames depicting touch occurrence and detection scenario. Starting from [b] showing the ring finger as moving fastest, then [c] shows the actual touch event and [e] showing the point of touch. Blue dot represents all tips extracted. Green dot represents fastest tip in that particular frame. Red dot represents actual point of touch. So, here a frame delay of two frames takes place to find the actual point of touch.



[e]

Fig. 5. [a]-[e]. (Continued)

This speed data is used to find fastest moving tip in forward direction in each frame. This data is maintained along with an additional information depicting for how many frames did the fastest tip moved in forward direction and then in backward direction. When their count is above a certain number of frames, the point of change of direction is considered to be as point of touch.

This point of touch is then finally traced to the key at this location.

3 Result and Discussion

The setup uses a single camera to extract the keys and watch the finger movements and conclusively detect touch events and map them to key locations. The camera used is i-ball super-view at the resolution of 640x480 but it can work for every camera with support for this resolution.

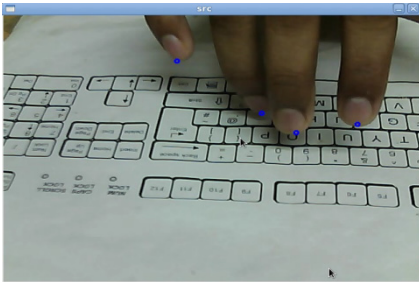
The overall procedure introduced here is new as a concept and provides handful of positive results. The idea of using the natural phenomenon that the finger being used to type moves fastest among all in previous frames works good enough to provide positive results.

The key localization module gives an accuracy of 100% in ideal computer generated keyboard images and up to 98% for those printed on paper or hand drawn on any surface but not so complex background. Also the surface doesn't needs to be flat for it to work and keys can be in any order. As visible in fig. 2 it works efficiently for several conditions as mentioned above.

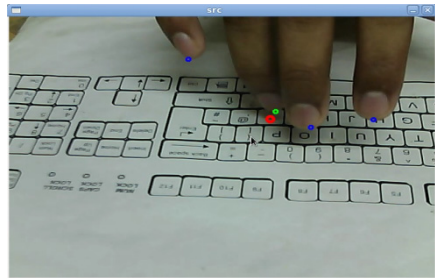
The hand segmentation module is quite effective in limited conditions specifically when surroundings do not have colour similar to that of skin. Tip extraction method works very accurate and is independent of object extracted in hand segmentation module. Thus if in case hand segmentation module is converted into segmentation of moving objects with tips, say pen, rather than being specific to hands, then it shall detect its tips.

The key touch detection method tends to have some improvements over other virtual keyboard methods that have been discussed earlier in the introduction part, namely, the method involving fuzzy rules for the hand gesture, machine learning based approach and the finger tip shadow method.

Specifically for the shadow based method consider a case as shown in following figure (Fig. 6). It shows index finger's shadow is overtaken by that of middle finger. This shall result into faulty touch detection. However, our method clearly differentiates between the two tips and thus being independent of the shadow shall provide better result.



[a]



[b]

Fig. 6. Situation in which shadow of one finger gets behind another finger

4 Conclusion

The method discussed here introduces a new paradigm of possibilities with virtual keyboards with fully customizable keyboard support and real time keystroke detection. However, many improvements can be introduced to it like using better algorithms for hand segmentation so as to make it more robust against various environments as the current one might not help with light source from an angle, dim light source, objects having similar colour to the skin like wood, etc.

Since the key identification module implemented was just for English, its working across other languages could not be verified but certainly the availability of several robust algorithms for multilingual optical character recognition [15] would help it get across this problem and help it become portable and accessible to everyone.

Another limitation this work has is that as of now it doesn't supports multi-touch functionality of the keyboards (e.g. ctrl+del, etc).

Acknowledgment. I sincerely would like to thank everyone who helped making this project a reality. Faculty members and my colleagues whose advices helped me improve it. Also would like to thank friends and family members who supported me throughout. At last, I would like to thank IIT-Allahabad authority for providing all necessary facilities and equipments.

References

1. Habib, H.A., Mufti, M.: Real Time Mono Vision Gesture Based Virtual Keyboard System. *IEEE 1266 Transactions on Consumer Electronics* 52(4) (November 2006)
2. Murase, T., Moteki, A., Suzuki, G., Nakai, T., Hara, N., Matsuda, T.: Gesture Keyboard with a machine learning requiring only one camera. In: *AH 2012 Proceedings of the 3rd Augmented Human International Conference*, Geneva (2012), doi:10.1145/2160125.2160154
3. Adajania, Y., Gosalia, J., Kanade, A., Mehta, H., Shekar, P.N.: Virtual Keyboard Using Shadow Analysis. In: *Third International Conference on Emerging Trends in Engineering and Technology*, November 19-21, 2011, pp. 163–165 (2011), doi:10.1109/ICETET.2010.115
4. ZhiYu, C., Abidi, B.R., Page, D.L., Abidi, M.A.: Gray-Level grouping (GLG): an automatic method for optimized contrast enhancement-part I: the basic method. *IEEE Transactions on Image Processing* 15(8), 2290–2302 (2006)
5. Kim, J., Han, Y., Hahn, H.: Character segmentation method for a License plate with topological transform. *World Academy of Science, Engineering and Technology* (2009)
6. Gupta, A., Srivastava, M., Mahanta, C.: Offline handwritten character recognition using neural network. In: *2011 IEEE International Conference on Computer Applications and Industrial Electronics (ICCAIE)*, December 4-7, pp. 102–107 (2011)
7. Haritaoglu, I., Davis, L.S., Harwood, D.: W⁴: Who? When? Where? What? A real time system for detecting and tracking people. In: *Third IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 222–227 (1998)
8. Wren, C., Azarbayehani, A., Darrell, T., Pentland, A.: PFinder: Real time tracking of human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19(7), 780–785 (1997)
9. Barron, J., Fleet, D., Beauchemin, S.: Performance of optical flow techniques. *International Journal of Computer Vision* 12(1), 42–77 (1994)
10. Gupta, R.K.: Comparative analysis of segmentation algorithms for Hand gesture recognition. In: *Third International Conference on Computational Intelligence, Communication Systems and Networks*, July 26-28, pp. 231–235 (2011)
11. Spruyt, V., Ledda, A., Geerts, S.: Real-time multi-colorspace hand segmentation. In: *17th IEEE International Conference on Image Processing*, pp. 3117–3120 (September 2010)
12. Bilal, S., Akmeliawati, R., Salami, M.J.E., Shafie, A.A., Bouhabba, E.M.: A hybrid method using haar-like and skin-color algorithm for hand posture detection, recognition and tracking. In: *International Conference on Mechatronics and Automation*, pp. 934–939 (August 2010)
13. Gang-Zeng, M., Yi-Leh, W., Maw-Kae, H.: Real-time hand detection and tracking against complex background. In: *Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pp. 905–908 (September 2009)
14. Oliveira, V., Conci, A.: Skin detection using HSV Color space. In: *Pedrini, H., de Carvalho, J.M. (eds.) Workshops of Sibgrapi 2009 –Posters*, pp. 1–2. SBC, Rio De Janeiro (2009)
15. Fujiyoshi, H., Lipton, A.J.: Real-time human motion analysis by image skeletonization. In: *Fourth IEEE Workshop on Applications of Computer Vision*, pp. 15–21 (October 1998)
16. Kae, A., Smith, D.A., Miller, E.: Learning on the fly: A font free approach towards multilingual OCR. *International Journal on Document Analysis and Recognition* 14(3), 289–301 (2011)

Human Factor Research of User Interface for 3D Display

Cih-Hung Ting¹, Teng-Yao Tsai¹, Yi-Pai Huang²,
Wen-Jun Zeng³, and Ming-Hui Lin³

¹Department of Photonics & Institute of Electro-Optical Engineering,

²Display Institute

National Chiao Tung University, Hsinchu, Taiwan, R.O.C. 30010

³Industrial Technology Research Institute, Hsinchu, Taiwan, R.O.C. 30010

chtting.eo98g@g2nctu.edu.tw

Abstract. The user interface for the observer to interact with 3D image has been discussed. The appropriate touching range and suitable size of the 3D image are relative to depth (disparity) of the 3D image. According to experimental results, when disparity of the 3D image is large, size of the 3D image is necessary to be larger to let the observer precisely judge that finger tip is touching the 3D image or not.

Keywords: User Interface, Interaction with 3D image, Appropriate Touching Range, Suitable Size of 3D Image.

1 Introduction

More intuitive and natural human-machine user interface (UI) is a tendency, such as from keyboard and mouse to touch panel. It is convenient and easy for users to interact with computers by using finger or stylus. However, general touch panel, smart phone or tablet, can only provide 2D image for user, and the user only makes a 2D interaction on the panel. On the other hand, although three-dimensional (3D) interaction is achieved by using camera or embedded optical sensor to catch user's position or movement, the user still watches 2D image without depth information [1]. Obviously, 3D display provides more realistic experience for observers, so 3D display has been more and more popular in many applications recently. Thus, the next step human-machine user interface should be 3D display with 3D interaction; that is, users can touch and interact with 3D image they watch, as shown in Fig.1. Nevertheless, there are maybe some issues for the user interface of 3D display as the observer makes 3D interaction with the 3D image.

Most of 3D display technology provides 3D image for observer by using binocular parallax, which means that the 3D display shows different image to left and right eye of the observer individually [2] [3]. However, the discrepancy between accommodation and convergence causes visual stress and leads to visual fatigue for observers [4]. There is still cross-link between accommodation and convergence when one eye of observer is occluded [5]. On the other hand, it must be blocked some part of 3D image by use's hand when the user wants to touch or interact with the 3D image. With

blocking some area of the 3D image, it may result in more serious mismatch or unstable for accommodation and convergence to cause some visual issues or mistakes. For example, the observer is more difficult to fuse those two images as mentioned above to create a 3D image, or even affects the observer's judgment on that finger tip is touching the 3D image or not. Therefore, a series of human factor experiments have been done in order to provide designing reference for the user interface as 3D image with different disparity, including the appropriate touching range and suitable size of 3D image.



Fig. 1. A schematic plot of 3D interaction with 3D image

2 Methods

2.1 Apparatus

An Acer 15.6 inch 3D notebook with pattern retarder mode was used to provide the 3D button for the subject. A prosthetic hand (or called finger tip) was used to be the subject's hand in following experiments, because sizes of subjects' hands were different and it was difficult to fix the subject's hand in the same position during experiment. The experimental parameters are shown in Fig. 2. Viewing distance of subject was 60 cm in front of the 3D notebook, and the button depth (disparity) was 0 cm and 0° when 3D button was displayed on the 3D notebook.

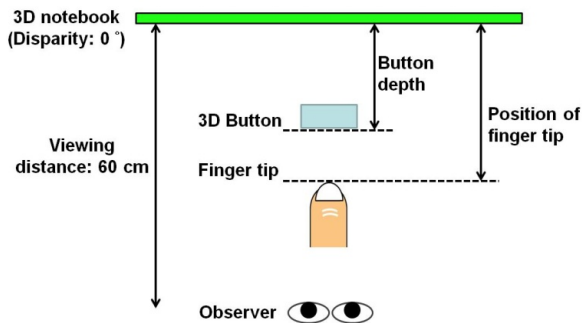


Fig. 2. Experimental parameters

2.2 Stimuli

Two kinds of button depth (disparity) were presented, 4.5 and 8.6 cm (disparities: 0.5° and 1°) with the interpupillary distance of the subject 65mm to prevent much visual discomfort [6] [7]. And in each session, distance between 3D notebook and finger tip was changed from 1~7 (button depth: 4.5 cm) and 5~11 (button depth: 8.6 cm) cm separately. In addition, blocking ratio was estimated by blocking area with finger tip divided by area of the 3D button without blocking in viewing position. There were four blocking ratios in each session, 20%, 40%, 60%, and 80%.

2.3 Subjects

Seventeen subjects, who had normal, uncorrected vision or wear optical correction to be corrected-to-normal vision, participated in this experiment. Subjects' ages were from 16 to 24 years. All subjects had normal stereoscopic vision and were unaware of the experimental hypotheses.

2.4 Experimental Setup

The device for measuring subject's perceived depth of the 3D button and experimental setup are shown in Fig.3. Before doing experiment, subject used the handle to align the blue sheet with the position where subject felt the 3D button was, in other words, subject's perceived depth of the 3D button. Additionally, the subject was fixed at the chin bracket to make sure that the position and viewing distance of different subjects were similar. Finally, the finger tip was moved by the mobile stage.

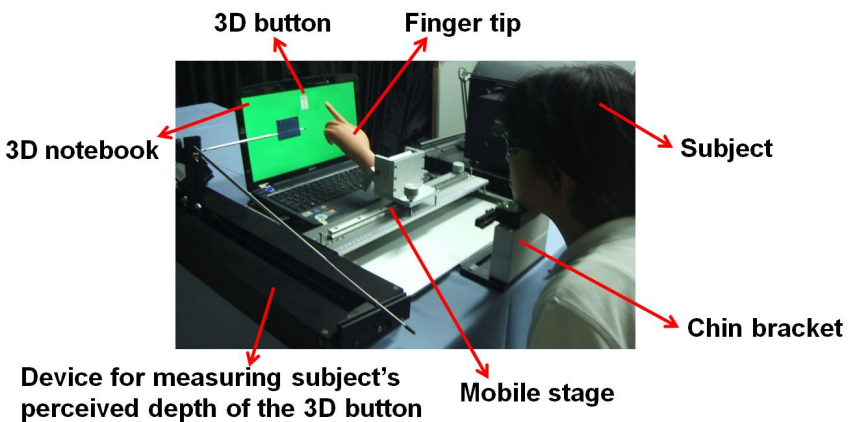


Fig. 3. Experimental setup

2.5 Procedure

First, the subject was instructed to close eyes to avoid seeing movement of finger tip. The finger tip was put randomly within 1~7 (button depth: 4.5 cm) and 5~11 (button depth: 8.6 cm) cm depending on which session. Second, the subject opened eyes, and selected which situation between 3D button and finger tip he perceived, over-touching, touching, or non-touching, as shown in Fig.4. In some cases, the subject felt that the button became a 2D button without depth or even could not fuse left and right eye image to be a 3D image; those two cases mentioned above were counted as non-touching.

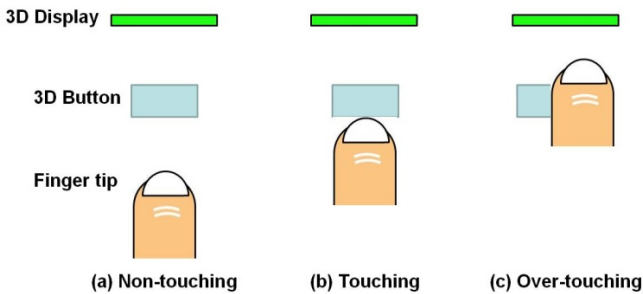


Fig. 4. Three kinds of situation between 3D button and finger tip

3 Results

3.1 Button Depth: 4.5cm (disparity: 0.5°)

As mentioned earlier, subjects were asked for measuring perceived depth of the 3D button before doing experiment. The average subjects' perceived depth was 4.4 cm; it was similar to button depth, 4.5 cm. On the other hand, with blocking by finger tip, the average subjects' perceived depth became 4.0 cm, which meant that blocking by finger tip reduced a little the subjects' perceived depth of the 3D button, as shown in Table 1. Further, the experimental result of subjects selecting different situations between 3D button and finger tip is shown in Fig.5. The appropriate touching range was 3~4 cm, for percentage of subjects selecting touching were 90% and 96%. Subjects also could judge over-touching and non-touching situation clearly. For instance, when finger tip was removed away from the 3D notebook to 6~7 cm, percentage of subjects selecting non-touching were 90% and 100%. Besides, regardless of blocking ratio, the average percentages of subjects selecting touching were similar, as shown in Table 2. It meant that blocking ratio did not influence subjects' judgment as 3D button with small button depth and disparity.

Table 1. The average subjects' perceived depth for button depth of 4.5 cm

Average subjects' perceived depth of the 3D button (Button depth: 4.5cm, disparity: 0.5°)	
Without blocking	With blocking
4.4	4.0

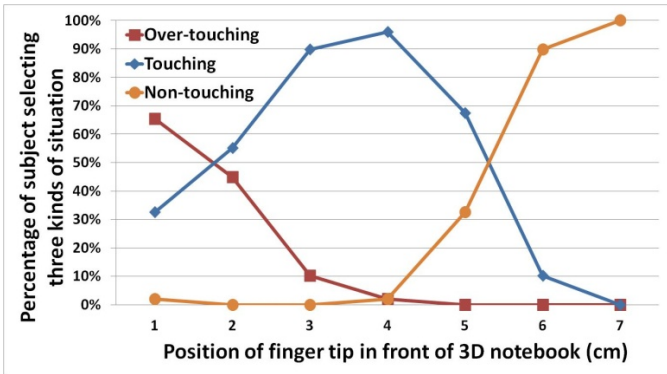


Fig. 5. Experimental result for subjects selecting different situations with different position of finger tip as depth of 3D button was 4.5 cm

Table 2. Average percentages of subjects selecting touching with different blocking ratio

Button depth: 4.5cm, disparity: 0.5°				
Blocking Ratio	20%	40%	60%	80%
The average percentage of subjects selecting touching within appropriate touching range	97%	90%	90%	100%

3.2 Button Depth: 8.6 cm (disparity: 1°)

The average subjects' perceived depth was 8.2 cm; it was still similar to button depth, 8.6 cm. However, with blocking by finger tip, the average subjects' perceived depth became 7.1 cm, as shown in Table 3. It meant that blocking by finger tip reduced more seriously the subjects' perceived depth of the 3D button than that of button depth of 4.5cm. The experimental result of subjects selecting different situations between 3D button and finger tip is shown in Fig.6. The appropriate touching range was 7~8 cm, for percentage of subjects selecting touching were 83% and 92%. Comparing Fig.6 with Fig.5, subjects' accuracy of judgment was lower not only within the appropriate touching range, but also without it. For instance, when finger tip was removed away from the 3D notebook to 10~11 cm, percentage of subjects selecting non-touching were only 47% and 62%. Moreover, according to Table 4, the average percentage of subjects selecting touching was decreased when blocking ratio was

increased. It might result from convergence being hard to fuse as the 3D button with large disparity and mismatch between accommodation and convergence was more unstable [8], so subjects' accuracy of judgment was reduced and subjects were necessary to have more information to judge that finger tip was touching the 3D image or not.

Table 3. The average subjects' perceived depth for button depth of 8.6 cm.

Average subjects' perceived depth of the 3D button (Button depth: 8.6cm, disparity: 1°)	
Without blocking	With blocking
8.2	7.1

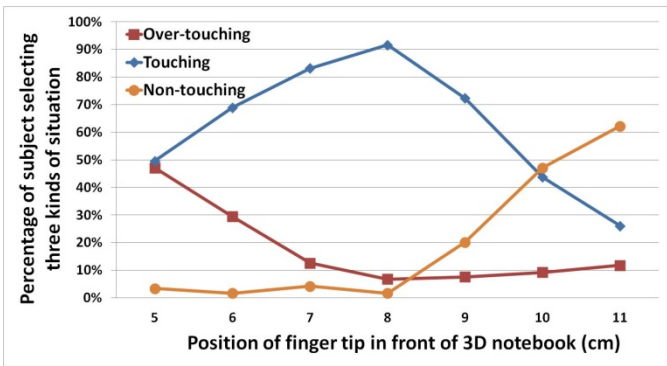


Fig. 6. Experimental result for subjects selecting different situations with different position of finger tip as depth of 3D button was 8.6 cm.

Table 4. Average percentages of subjects selecting touching with different blocking ratio.

Button depth: 8.6cm, disparity: 1°				
Blocking Ratio	20%	40%	60%	80%
The average percentage of subjects selecting touching within appropriate touching range	91%	88%	88%	76%

4 Conclusion

A series of human factor experiments have been done in order to provide designing reference for the user interface of 3D display as 3D image with different disparity, including appropriate touching range and suitable size of 3D image. According to experimental results, regardless of disparity of 3D image, the appropriate touching range is both about 1 cm, but blocking by the finger tip reduces more seriously the

subjects' perceived depth of 3D image with large disparity than that with small disparity. Besides, when disparity of 3D image is large, size of 3D image is necessary to be large to let the observer precisely judge that finger tip is touching the 3D image or not because they need more information. In conclusion, if observers want to interact or touch the 3D image with large disparity, size of the 3D image should be larger than that with small disparity.

Acknowledgement. This research was supported by Industrial Technology Research Institute (ITRI) of Taiwan.

References

1. Wang, G.-Z., et al.: A Virtual Touched 3D Interactive Display with Embedded Optical Sensor Array for 5-axis (x, y, z, θ, φ) Detection. In: SID Symposium Digest (2011)
2. Yoshihara, Y., Ujike, H., Tanabe, T.: 3D crosstalk of Stereoscopic (3D) display using Patterned Retarder and Corresponding Glasses. IDW Digest (2008)
3. Urey, H., Chellappan, K.V., Erden, E., Surman, P.: State of the Art in Stereoscopic and Autostereoscopic Displays. Proceedings of the IEEE 99(4) (April 2011)
4. Hoffman, D.M., Girshick, A.R., Akeley, K., Banks, M.S.: Vergence–accommodation conflicts hinder visual. *Journal of Vision* 8(3), 33, 1–30 (2008)
5. Ukai, K., Howarth, P.A.: Visual fatigue caused by viewing stereoscopic motion images: Background, theories, and observations. *Displays* 29, 106–116 (2008)
6. Iwasaki, T., Kubota, T., Tawara, A.: The tolerance range of binocular disparity on a 3D display based on the physiological characteristics of ocular accommodation. *Displays* 30, 44–48 (2009)
7. Kooi, F.L., Toet, A.: Visual comfort of binocular and 3D displays. *Displays* 25, 99–108 (2004)
8. Ukai, K., Kato, Y.: The use of video refraction to measure the dynamic properties of the near triad in observers of a 3-D display. *Ophthalmic and Physiological Optics* 22(5), 385–388 (2002)

Collaborative Smart Virtual Keyboard with Word Predicting Function

Chau Thai Truong¹, Duy-Hung Nguyen-Huynh¹,
Minh-Triet Tran¹, and Anh-Duc Duong²

¹ University of Science, VNU-HCM, Vietnam

² University of Information Technology, VNU-HCM, Vietnam
{0912034, 0912202}@student.hcmus.edu.vn,
tmtriet@fit.hcmus.edu.vn, ducda@uit.edu.vn

Abstract. The authors propose a table-top with virtual keyboards for multi-users to work in a collaborative environment. The proposed system has two main modules: a system for virtual keyboards with touch event detection from depth data of a Kinect and a word predicting module based on the idea of Hidden Markov Model and Trie data structure. The system can replace physical keyboards, improve the accuracy of a virtual keyboard, and increase the typing speed of users. Our experimental results show that our system archives an accuracy of 94.416% with the virtual keyboard, saves 11-22% of keystrokes, and corrects 89.02% of typing mistakes.

Keywords: table top, virtual keyboard, word prediction, 3D interaction.

1 Introduction

Human-computer interaction (HCI) plays an important role in the evolution of computing society. Researches in the field of interaction between users and computers aim to enhance the comfort, ergonomics, and portability as well as to save time for users. To efficiently assist users in various activities in daily life, HCI not only provides useful utilities for single users to interact with computing systems but also collaboration environment for multiple users to work together.

Text-based data entry is one of the most common tasks in most applications. Thus different types of keyboards have been developed to enhance the usefulness and comfort for users. Inspired by the augmenting interactive table system with mice and keyboards of Hartmann [6], we propose a collaborative smart virtual keyboard system with word predicting function. Our proposed system uses a regular projector to project over an arbitrary relatively-flat surface the images of multiple virtual keyboards and a Kinect to capture depth information for touch event detection. Multiple users can now work together on a single large area, e.g. a desk, using only virtual devices. Furthermore, the layout, language, and size of any virtual keyboards can be visually customized to save extra cost for real physical devices to meet users' various needs.

Besides, the word predicting function is designed to support users to increase the typing speed and correct typing mistakes quickly. We use a dictionary stored in a

prefix tree (Trie [12]) to save memory and to increase processing speed. The dictionary contains both words and their frequencies in the same category, e.g. sports, education, politics, etc. There are two parts of this function: predicting the words that a user intends to type and correcting words that a user mistyped. The operations of these parts are based on word frequencies in a trained dictionary and the prefix characters that a user has just typed. Experimental results show that our virtual keyboard system achieves the accuracy of 94.42% in keystrokes detection, the word prediction can save 10-20% keystrokes, and the word correction can eliminate up to 80-90% mistakes when users type text data with the same topic as that of the used dictionary.

The content of the paper is as follows. In Section 2, the authors briefly review the approaches in HCI, especially in developing table top and virtual keyboards. Our proposed system and experimental results are presented in Section 3 and 4 respectively. Conclusions are discussed in Section 5.

2 Related Work

Interactive surfaces of different materials, technologies, and sizes have become popular means of interaction between users with computing devices and systems. While touch screens are suitable for small and medium sized devices, such as tablets, mobile devices, ATM machines, etc., tabletops are more applicable for economic and large sized interactive surfaces.

Different approaches have been proposed to develop various models of tabletops, such as using laser to detect and localize touch events [1], using single or multiple traditional cameras to detect hands and fingers' actions with or without markers/gloves [2], using sensors of multi-touch surfaces to perform interaction [10]. With the appearance of depth-sensing technology and depth cameras such as Kinects, vision-based methods for tabletop interaction can take a further step with extra useful information of depth data. Wilson uses depth data as touch sensor and can determine interaction points by using a depth camera [3]. In this paper, we follow this new trend to develop our smart virtual keyboard with depth data captured from a Kinect.

Among with tabletops, applications that support multi-user work or collaboration are also a topic of concern. A multi-user web browser system is proposed to support multiple people to search and to watch the same webpages simultaneously [8]. Klinkhammer et. al. develop a system that enables many people to share information, data while working on the same interactive surface [9]. WeSearch system is proposed to enable a group of up to four members to use a web browser at the same time on a tabletop [7]. Especially, Hartmann et. al. propose eight interaction methods that are used in a working desk that supports interactions with real keyboards and mice [6]. This motivates our inspiration to develop our system. However, we take a further step. Our system does not require any physical input devices, e.g. keyboards, mice. Users interact through virtual devices projected by a projector over any relatively-flat surface and touch events are detected from depth data captured from a Kinect. Therefore, keyboard layouts and languages can be customized easily.

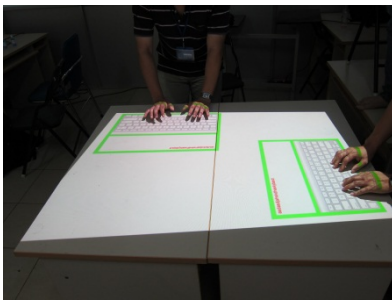
Beside applications of tabletops and supports for collaboration, the efficiency in typing on virtual keyboards is also a practical demand. Findlater et. al. propose a

method to evaluate an adaptive personalized virtual keyboard layout to improve typing on a touch-screen [4]. Wigdor et. al. examine the hand patterns in touch-typing on a flat surface to suggest a new design for touch screen keyboards [5]. In this paper, our approach is to perform word prediction and correction using word frequencies in a dictionary containing words in the same category. This approach has many advantages. Different dictionaries can be built by training large data from online articles of the same topic from the Internet. The frequencies of words can be combined with the neighborhood keys to correct mistyped words.

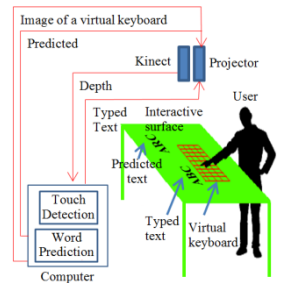
3 Proposed System

This section shows the main components and operations of our smart virtual keyboard with word prediction and correction. The overview of our system is presented in Sec. 3.1. The operations of Kinect are described in Sec.3.2. Finally, Sec.3.3 shows the operations of word prediction function including the data structure to store a dictionary and our proposed methods to predict words from prefixes.

3.1 Overview of the System



(a) Overview



(b) Main components

Fig. 1. Overview and architecture of the system

Figure 1 illustrates the overview and main components of our proposed system. A user interacts with a virtual keyboard displayed on a non-touch interactive surface. In our system, the interactive surface and the Kinect are in fixed positions during an interaction process. A projector is used to project images of multiple virtual keyboards, texts, and multimedia information on demand. A single computer is used to processed all keystrokes from multiple users and other functions. These functions are divided into two subcomponents: touch detection and word prediction.

Touch detection: A Kinect device is used to continuously capture depth images on the interactive surface. The computer receives depth information to detect touched points on virtual keyboards and generates appropriate keyboard events. Finally, the projector shows the image of virtual keyboards and typed texts. The role of Kinect in this component is described in detail in Sec.3.2.

Word prediction: this subcomponent is designed to enhance the accuracy of touch detection and to increase typing speed. When a prefix is typed, the system suggests appropriate words to the user. The use of data structure to store a dictionary and our proposed methods to predict words are presented in Sec.3.3.

3.2 Touch Event Detection with a Kinect

In this part, we present the use of a Kinect in touch detection. First, a background image is trained to estimate the main interaction plane. It should be noticed that the interaction plane is not required to be perfectly flat. Any fixed, relatively flat surface can be used as the interaction plane.

Then, when users start using the system, query images of depth data are computed periodically to detect touched points. However, when computing the background image and query images, due to the instability of Kinect depth data, the depth information in an image should be denoised on consecutive frames by a median filter.

Background estimation: A background image contains only depth data of the interactive surface. Therefore, the computation of the background image must be done on many continuous frames at the beginning of an interaction process. Let N_1 be the number of depth frames used in training the background, $I(x, y)$ be the value of a pixel (x, y) in an image I , I_B be the background image and $I_{B_1}, I_{B_2}, \dots, I_{B_{N_1}}$ be the frames that are used to compute I_B . We estimate the background image with the following formula: $I_B(x, y) = \text{median}\{I_{B_i}(x, y); i \in [1, N_1]\}$

Computing query images: Suppose the process starts at the s -th frame. Let N_2 be the number of consecutive depth frames used to compute a query image. The i -th query image ($i > 0$), denoted by I_Q is computed from the frame $s + (i - 1) \cdot N_2$ to the frame $s + i \cdot N_2 - 1$ by the following formula:

$$I_Q(x, y) = \text{median}\{I_{Q_i}(x, y); i \in [s + (i - 1) \cdot N_2, s + i \cdot N_2 - 1]\}$$

Finding the points that are near the interactive surface: Depending on its real distance D_p to Kinect, a point P in a depth image is classified into one of the three classes: inside, near, and far point. The depth value at a pixel is the distance from Kinect to the plane that contains the pixel and is perpendicular to the viewing direction of Kinect. Let D' be the distance from Kinect to the plane containing P . $D' > D_p$ means that P is inside the interactive surface. This is certainly a noise data because Kinect and the surface are assumed to be stable. Therefore, if $D' > D_p$, we reset $D' = D_p$ to eliminate noise. Finally, we have $D' \leq D_p, \forall P \in I_Q$.

Let $\Delta_p = D_p - D'$, $\Delta_{\min} = \min\{\Delta_p \mid \forall P \in I_Q\}$ and $\Delta_{\max} = \max\{\Delta_p \mid \forall P \in I_Q\}$.

We define the set of near points $C = \{P \in I_Q \mid \Delta_{\min} \leq \Delta_p \leq \Delta_{\max}\}$. With the same value of Δ_{\max} , the smaller value of Δ_{\min} causes more noise data between Δ_{\min} and Δ_{\max} . Otherwise, the larger value of Δ_{\min} makes the touch events generated earlier. With the same value of Δ_{\min} , if Δ_{\max} is larger, the noise between Δ_{\min} and Δ_{\max} is higher. Otherwise, the time interval of a touch event is shorter and a touch event is more likely to be missed. The experiment to choose Δ_{\min} and Δ_{\max} is presented in Sec.4.1.

Touch detection: For a set T containing connected near points, if $|T| > \min_{Area}$, a touch event is triggered at the position (x_{touch}, y_{touch}) where:

$$x_{touch} = \frac{\sum_{p \in T} P.x}{|T|} \quad \text{and} \quad y_{touch} = \frac{\sum_{p \in T} P.y}{|T|}$$

3.3 Word Prediction and Correction

When a user types a prefix, i.e. a sequence of initial characters of a word, the system suggests a number of words with highest probabilities. If the user accepts, he or she can save time to type that word. This feature is also used to correct typing mistakes. These mistakes can be typing mistakes of users and touch event mistakes caused by the system. When a prefix is typed, the system corrects and suggests new words based on the neighborhood characters of the typed characters. Using this feature, users can save time for both typing and correction.

Both functions use dictionaries to perform. The dictionary contains frequencies of words in the same topic with the typing text. Together with the dictionary, user profiles can also be utilized to enhance the accuracy of these functions. Therefore, the more often a user uses the system, the more efficient the system becomes.

Prefix tree (Trie): Trie data structure is first proposed by Fredkin [12]. Trie is a tree $G = (V, E)$ with a root r , a vertex set V , and an directed edge set $E = (V, V)$. Each edge $(u, v) \in E$ contains a character $c_{(u,v)}$. For a node x and its children nodes y_i and y_j , we have $c_{(x,y_i)} \neq c_{(x,y_j)}$ for $y_i \neq y_j$. Figure 2 illustrates an example of a trie.

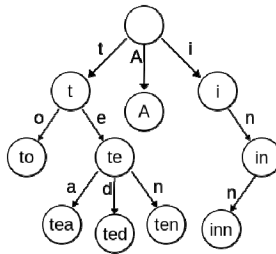


Fig. 2. Trie built from strings “A”, “to”, “tea”, “ted”, “ten”, “i”, “in” and “inn”

The efficiency of using Trie in dictionary: Let S be the set of words in a dictionary, $\text{len}(x)$ be the length of string x and $|\overline{u_i}|$ be the average length of all words in the dictionary, we consider the following operations.

- Storing a set of string $T = \{v_i\}$ with the same prefix u .

Assume that we need one memory unit to store a character, the number of memory units that can be saved when using Trie to store T is as follow:

Total number of characters in T – number of memory unit to store using Trie

$$\begin{aligned}
&= \left(\sum_{v_i \in T} (\text{len}(v_i)) \right) - \left(\text{len}(u) + \sum_{v_i \in T} (\text{len}(v_i) - \text{len}(u)) \right) \\
&= (|T| - 1) \cdot \text{len}(u)
\end{aligned}$$

- Finding a prefix $v = v_1v_2\dots v_k$ in S

When not using Trie, the system iterates through every word or binary searches the wordlist to find v . The complexity of these operations are $O(|S| \cdot |\overline{u_1}|)$ and $O(\log_2 |S| \cdot |\overline{u_1}|)$ respectively. When using Trie, the system starts from root r and visits child nodes until v is found or cannot visit more. The complexity is then $O(|\overline{u_1}|)$. Therefore, using Trie can reduce searching time of a prefix.

- Adding a new word $v = v_1v_2\dots v_k$ to S

When not using Trie, the complexity to insert v at position i in alphabetical order is $O(|S| |\overline{u_1}|)$. When using Trie, the system finds a prefix of v . Then, it iterates through every remaining position to add new child nodes. The complexity of this is $O(|\overline{u_1}|)$. Therefore, the adding time is faster.

In conclusion, the use of Trie to store dictionaries has three advantages: smaller capacity, faster time to search and add words. That is the reason of our choice to use Trie to perform main functions for word prediction and correction in our system.

Hidden Markov Model (HMM) [11] has the following elements:

- A finite number, denoted by N , of states in the model. At each clock time, t , a new state is entered based upon a transition probability distribution which depends on the previous state.
- After each transition is made, an observation output symbol is produced according to a probability distribution which depends on the current state. This probability distribution is held fixed for the state regardless of when and how the state is entered. There are thus N such observation probability distributions which, of course, represent random variables or stochastic processes.

All of those elements are formally defined in the following:

T = length of the observation sequence

N = number of states in the model

M = number of observations

$Q = \{q_1, q_2, \dots, q_N\}$, states

$V = \{V_1, V_2, \dots, V_M\}$, observations

$A = \{a_{ij}\}$, $a_{ij} = \Pr(q_j \text{ at } t + 1 \mid q_i \text{ at } t)$, state transition probability distribution

$B = \{b_j(k)\}$, $b_j(k) = \Pr(v_k \text{ at } t \mid q_j \text{ at } t)$, observation probability distribution in state j

$\pi = \{\pi_i\}$, $\pi_i = \Pr(q_i \text{ at } t = 1)$, initial state distribution

Prediction based on frequencies of prefixes: When a user types a prefix s , a complete word u_i that has the prefix s and has the highest probability is suggested. Let sum_s be the frequency of s . The probability of u_i is defined as follow:

$$P(u_i | s) = \frac{sum_{u_i}}{sum_s}$$

Prediction based on finger position and keyboard layout: Let $s = s_1s_2...s_m$ be the typed prefix, the system suggests the list of prefixes $u = u_1u_2...u_m$ that has the highest probability. The suggestions of prefixes u depends on two criteria:

- The frequency of u in a dictionary.
- The neighborhood characters of s_i (for $i \leq m$) is the set $H(s_i)$ of characters b such that the keys of s_i and b are identical or share a common part in key borders. For example, in a standard US keyboard, the keys E, R, F, C, X, S are neighborhoods of D. These keys b have the nearest distances from their centers to the center of key s_i . Therefore, we choose this set as a criterion. Let $P_t(s_i, b)$ be the probability that a user actually wants to type b instead of s_i . We assign $P_t(s_i, b) = r\%$ when $s_i \neq b$. $P_t(s_i, s_i)$ changes depending on r and $|H(s_i)|$.

Let $\sigma(u, i) = u_1u_2...u_i$. The application of hidden Markov model is as follows:

- The states are the prefixes $\sigma(u, i)$ ($i \leq m$) that exist in the dictionary and u_i is a neighborhood of s_i .
- The observations of the current state $\sigma(u, i)$ ($i < m$) is the set $H(s_{i+1})$.
- Transition probability distribution A : if $i < j$, $\sigma(u, i)$ is a prefix of $\sigma(u, j)$. We have:

$$A_{\sigma(u,i),\sigma(u,j)} = \frac{\text{sum}_{\sigma(u,i)}^{\sigma(u,j)}}{\text{sum}_{\sigma(u,i)}}$$

- Observation probability distribution B : given the state $u' = \sigma(u, i)$ ($i < m$) and $H(s_{i+1})$ are the observations. We have: $B_{u',c} = P_t(s_{i+1}, c)$ where $c \in H(s_{i+1})$
- Initial state distribution π_u : Let P be the set of all prefix in the dictionary, we have:

$$\pi_u = \frac{\text{sum}_u}{\sum_{t \in P} \text{sum}_t}$$

- The probability of an observation series $u_1u_2...u_m$ is:

$$\begin{aligned} + \text{ For } m = 1, \text{ we have } P(u_1) &= P(\emptyset \rightarrow u_1) * P_t(s_1, u_1) = A_{\emptyset, u_1} * B_{\emptyset, u_1} \\ &= \pi_{u_1} * P_t(s_1, u_1) \text{ where } \emptyset \text{ is the empty string.} \end{aligned}$$

+ For $m > 1$, we have:

$$\begin{aligned} P(u_1u_2 \dots u_m) &= \prod_{i=1}^{m-1} P(\sigma(u, i-1) \rightarrow \sigma(u, i)) * P(\sigma(u, i-1) \rightarrow u_i) \\ &= \prod_{i=1}^{m-1} A_{\sigma(u,i-1),\sigma(u,i)} * B_{\sigma(u,i-1), u_i} = \prod_{i=1}^m \frac{\text{sum}_{\sigma(u,i)}}{\text{sum}_{\sigma(u,i-1)}} * P_t(s_i, u_i) \end{aligned}$$

4 Experimental Results

In this section we present four experiments. The experiment in Sec.4.1 determines the optimum values of Δ_{\min} and Δ_{\max} for touch detection. Sec.4.2 measures the accuracy of our virtual keyboard with the chosen values of Δ_{\min} and Δ_{\max} in Sec.4.1. The experiments in Sec.4.3 and Sec.4.4 are to evaluate the efficiency of word prediction based on word frequencies and the accuracy of word correction based on both word frequencies and neighborhood characters. These experiments are performed on the system using CPU core i7 2.2Ghz, 6GB RAM.

4.1 Finding Optimal Values of Δ_{\min} and Δ_{\max}

This experiment is to find the optimum values of Δ_{\min} and Δ_{\max} to achieve the highest precision in touch detection. For experiments in this section and Sec.4.2, Kinect is kept at a distance of 0.8m from the surface. 26 square regions with the same size of 1.5cm \times 1.5cm are selected to be the virtual keys corresponding to 26 alphabet characters from 'a' to 'z'. The data consists of 1800 depth frames in 640 \times 480 resolution. The first 300 frames are used to train the (3D) background. Touching action is performed at arbitrary regions in the remaining frames. The accuracy is determined as the proportion between the number of error frames and the total number of used frames. From the experiment, we choose $\Delta_{\min} = 10\text{mm}$ and $\Delta_{\max} = 13\text{mm}$.

4.2 Estimating the Accuracy of Virtual Keyboard

The purpose of this experiment is to measure the accuracy of our virtual keyboard with the optimum parameters Δ_{\min} and Δ_{\max} determined in Sec.4.1. The data consists of 10 strings with 154 characters in total. For each string, we type 5 times to the virtual keyboard. The accuracy of each typing time is the percentage of correct touch events. After typing all of 10 strings, we calculate the average accuracy of five times over all strings. Table 1 shows the results.

Table 1. Accuracy of virtual keyboard in 5 typing times

Test	Result
1	94.156%
2	94.805%
3	92.208%
4	94.805%
5	96.104%
Average	94.416%

From Table 1, we conclude that our system archives the average accuracy of 94.416% in touch detection of our virtual keyboard.

4.3 Estimating the Efficiency of Word Prediction Based on Word Frequency

In this experiment, the efficiency of word prediction is measured by the percentage of keystrokes that users can save when they type texts in the same topic as that of the current dictionary. We measure two methods: offline – word frequencies are fixed; and online – the word frequencies are updated gradually after each time a user types a word. A dictionary containing words and their frequencies is built from 100 articles in a single topic (e.g. business, technology, health, etc). 100 other articles in the same topic are selected to test the system. Each article in the training set and test set has 2000-6000 words. The result is illustrated in Figure 3.

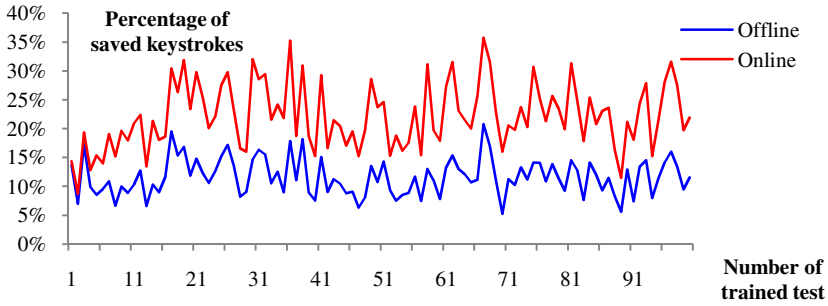


Fig. 3. Percentage of saved keystrokes when applying two methods (online and offline) of word prediction based on word frequencies

From the experimental results, we conclude that with the same text and topic, when the number of typed words is small, the accuracy of two methods do not have much difference. When the number of words increases, the efficiency of online method becomes higher. On average, users can save 11.61% keystrokes with offline method and 22.19% with online method.

4.4 Estimating the Accuracy of Word Correction Based on Word Frequency and Neighborhood Characters

This experiment aims to measure the accuracy of system to correct typing mistakes of users based on the frequencies of words in the same specific topic and the neighborhoods of the typed characters. Let P_{error} be the probability that a character is mistyped to a neighborhood character, we measure the percentage P_{fix} of corrected mistakes after using our method. A dictionary are built from 40 articles in a single topic and 40 other articles are used as testcases. Each article has 2000-6000 words and the total number of characters is 85627. Recall the Sec.3.3, we choose $P_t(a, b) = 5\%$ if $a \neq b$. The result of this experiment is in Figure 4.

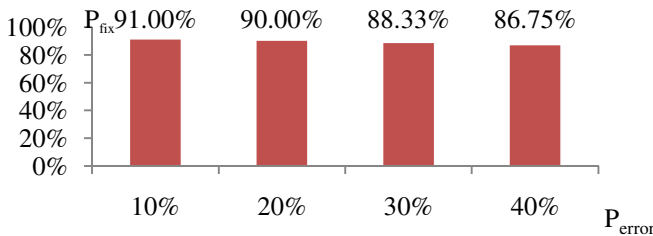


Fig. 4. Result of the efficiency of word correction using word frequencies and neighborhood characters

From the experimental result, we conclude that using word frequencies and neighborhood characters to correct typing mistakes can solve 89.02% on average of these mistakes when typing in the same topic. This percentage decreases when the initial error rate increases.

5 Conclusions

We propose to develop a table top with smart virtual keyboards for multiple users to work in a collaborative environment. To eliminate noise in Kinect, we use a median filter in both background training and touch event detection. The virtual keyboard system can be apply over any relatively flat area with the accuracy of 94.416%. The position, layout, and language of each virtual keyboard can be customized easily.

Besides, the accuracy of our virtual keyboard is enhanced by word predicting function that learns from both train data and user profile to suggest and correct the typed words quickly. This feature helps users to save up to 11-22% of keystrokes and correct 89.02% mistakes in typing documents in the same topic with the dictionary.

We are currently develop further features for our tabletop environment and improve the system to apply on any types of surfaces (curve, sphere, etc). Besides various dictionaries are being built for different topics and categories, even for programming languages (e.g. C++, Pascal, C#).

References

1. Tobias Schwirten, L.: Radar Touch, <http://www.radar-touch.com/>
2. Marquardt, N., Kiemer, J., Greenberg, S.: What caused that touch? expressive interaction with a surface through ducinary-tagged gloves. In: ACM International Conference on Interactive Tabletops and Surfaces, ITS 2010, pp. 139–142. ACM (2010)
3. Wilson, A.D.: Using a depth camera as a touch sensor. In: ACM International Conference on Interactive Tabletops and Surfaces, ITS 2010, pp. 69–72. ACM (2010)
4. Findlater, L., Wobbrock, J.: Personalized input: improving ten-finger touchscreen typing through automatic adaptation. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2012, pp. 815–824. ACM (2012)
5. Findlater, L., Wobbrock, J.O., Wigdor, D.: Typing on flat glass: examining ten-finger expert typing patterns on touch surfaces. In: Proceedings of the International Conference on Human Factors in Computing Systems, CHI 2011, pp. 2453–2462 (2011)
6. Hartmann, B., Morris, M.R., Benko, H., Wilson, A.D.: Augmenting interactive tables with mice & keyboards. In: Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology, UIST 2009, pp. 149–152. ACM (2009)
7. Morris, M.R., Lombardo, J., Wigdor, D.: Wesearch: supporting collaborative search and sensemaking on a tabletop display. In: Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work, CSCW 2010, pp. 401–410. ACM (2010)
8. Tuddenham, P., Davies, I., Robinson, P.: Websurface: An interface for co-located col-laborative information gathering. In: Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces, ITS 2009, pp. 181–188. ACM (2009)
9. Klinkhammer, D., Nitsche, M., Specht, M., Reiterer, H.: Adaptive personal territories for co-located tabletop interaction in a museum setting. In: Proc. of the ACM International Conference on Interactive Tabletops and Surfaces, ITS 2011, pp. 107–110. ACM (2011)
10. Dippon, A., Echtler, F., Klinker, G.: Multi-touch Table as Conventional Input Device. In: Stephanidis, C. (ed.) Posters, Part II, HCII 2011. CCIS, vol. 174, pp. 237–241. Springer, Heidelberg (2011)
11. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE, 257–286 (1989)
12. Fredkin, E.: Trie Memory. CACM 3(9), 490–499 (1960)

The Implementation of Multi-touch Table to Support the Military Decision Making through Critical Success Factors (CSFs)

Norshahriah Wahab and Halimah Badioze Zaman

Institute of Visual Informatics, Universiti Kebangsaan Malaysia
shahriah@upnm.edu.my, hali@ivi.ukm.my

Abstract. In this paper, we present the implementation of Multi-touch Table (MTT) to support the Military Decision Making. The need of the multi-touch table technology is essential for effective and efficient outcome especially in the Malaysian Environment Army. The decision making process is also integral to successful performance of the battlefield. The military decision making process emphasized on timely decision making, the understanding between commander's intent and staff besides the clear responsibility of the commander and staff. Therefore, the crux of this paper is on how to optimize the military decision making process through the Critical Success Factors (CSFs) that have been identified from preliminary study. By adapting the Critical Success Factors (CSFs), all the concepts, ideas and arguments can be brainstorm clearly and effectively around the Multi-touch Table which further gives advantages in visualizing, organizing and manipulating the data/information amongst military officers. The adaptation of the elements in Critical Success Factors (CSFs) also will promote the communication between commander and staff in the activities that involved visualizing the battle-space, describing the visualization to subordinates/staff, directing action in terms of the battlefield operating system and leading the unit to mission accomplishment. This paper also will present the findings and results obtained from series of questionnaires and interviews amongst Subject Matter Experts (SME) in the domain of Military Decision Making. Based on preliminary study indicated that the Criticality of elements in Critical Success Factors (CSFs) in supporting the process of military decision making. One big issue or dilemma in planning and execution of military decision making is the Commanding Officer (CO) need to rely fully on the subordinate officers' coordination ability and to understand effectively of the consequences each 'Course of Action' (COA) suggested by subordinates officers. The application of Multi-touch Table will be benefited in term of the medium used in supporting the discussion and brainstorming session between the Commanding Officer (CO) and the subordinate staff. Decision makers will refer to the shared display together at the same time with different orientations. Multi-touch Table is interactive table that becoming affordable in commonplaces such as in offices, universities and homes. This technology offers the world possibilities such as task engagement, face-to-face communication, social interaction dynamics and simultaneous input contribution. In the nut shell, the appropriate medium such as Multi-touch Table will put the positive impact

towards the process in military decision making and addition to this point the adaptation of Critical Success Factors (CSFs) may give a lot of advantages specifically in planning and execution of military decision making.

Keywords: Multi-Touch Table, Military Decision Making, Critical Success Factors (CSFs), Command and Control (C2).

1 Introduction

Decision making has been researched from a number of paradigms, including the classical decision making approach which advocates a logical, rational, analytical approach to decision making suited to military planning and naturalistic decision making which reflects decision making in uncertain and dynamic military operational environments (Daley, 2007). Decisions in a military context, is characterized as “command and control” (C2) decision making and this category features high-level integration of near real-time information for the purpose of deciding how best to utilize force application in a battle environment under varying degrees of uncertainty and time pressures (George, 2008).

Besides, the data and information need to be presented in the collaborative and effective interface. According to Hancock et.al (2006), the surface of Multi-touch Table gives advantages in terms of the vast, horizontal surface and multiple users may use it in the same time. The application of Multi-touch Table offers one type of mechanism that enables the users to change the orientation of the interface by using one touch or multiple fingers. By this way, the users can easily and freely rotate the display orientation that suit and appropriate to them. This can be seen in Figure 1.



Fig. 1. The application of Multi-touch Table by multiple users at the same time, different orientations

Furthermore, there are others research that have been proved the advantages of Multi-touch Table as a tool or technology which very helpful especially in communication and interaction. Based on research conducted by Cummings et.al (2011), stated

that the Multi-touch Table technology is effective and efficient tool for handling the situation that involved with a huge amount of data/information regardless displayed in one time and this data/information are the main focus of discussion. Moreover, research by Scott (2010), also proved that the application of Multi-touch Table will assist and support the process of military decision making in collaborative setting specifically.

Based on the facts, the implementation of Multi-touch Table will enhance the communication and interaction, the elements in Critical Success Factors (CSFs) in order to support the “command and control” process in military. By putting the communication and interaction as factors that contributing the success in military decision making, supporting with the level of Situational Awareness (SA) and Working experience, will promote the decision making amongst military staffs and Commanding Officer (CO). The elements of Critical Success Factor (CSFs) needed to be emphasized in order to optimize the decision making process. According to (Burns, 2000), situation/shared awareness as the construction of “mental models” to describe, explain and predict a situation. In particular, (Endsley, 2000) proposes that there are three (3) levels of situation/shared awareness as follows :

- i. “description” (of situational elements)
- ii. “explanation” (of the current situation)
- iii. “prediction” (of future states and actions)

Based on preliminary study that have been conducted, thirty (30) respondents from military peoples have answered to the questionnaire sample and the findings depicted that three (3) major factors currently most contributing to optimize the military decision making are Communication and Interaction (40%) followed by Shared Awareness (20%). This is depicted in Figure 2.0 as follows.

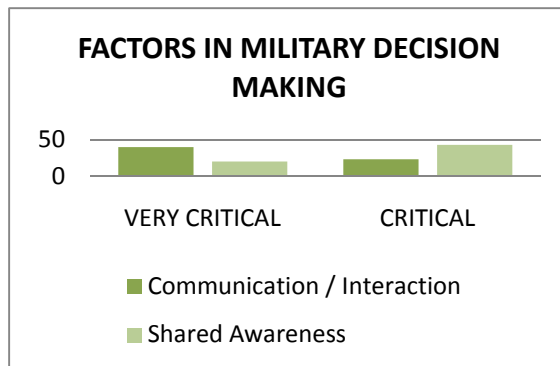


Fig. 2. Figure 2.0: Critical Success Factor (CSFs) in Military Decision Making

2 Preliminary Analysis

The preliminary analysis has been conducted in order to have the feedback and responses on the current problem in the planning and execution of military decision making process. The set of questionnaires have been disseminated amongst thirty (30) respondents involving captains at battalion level. The result can be depicted as follow in Table 1.0.

Table 1. Result from questionnaires' respondents from MINDEF, Kuala Lumpur

Item No.	Question and Answer/s	Percentage (%)
9G	Currently, is there any system/tools used to support your department in the process of decision making? If YES, please state.	Non (93.3%)
C 11	The following are the problems or constraints in the process of decision making. Please indicate how critical these problems are in affecting the decision maker, based on the given scale. (I3) : Communication between decision maker to staff (CO) (I6) : Situational Awareness (I7) : Experiences	Very Critical & Critical (40%) & (20%) (20%) & (44%) (23%) & (4%)
I3	According to list below, please identify the importance of the tools/applications in order to support the process of decision making based on the given scale. (K4) : 'Multi-touch Table/ Touch screen'.	(50%)

3 Findings of Preliminary Study

Based on the research outcomes from Table 1, the conclusions can be made as follows:

- i. There is (93.3%) responded that non of system/tool that currently used to support the process of military decision making. Otherwise (6.7%) stated that there is a system/tool used in their unit such as simulation to support their decision making process. This statement can clearly be seen in Figure 3.0 below.
- ii. The factors that contributing to Critical Success Factor (CSFs) are communication and interaction amongst military officers besides the importance of Situational Awareness (SA). Situational awareness is needed to achieve the mission goal. Moreover, the

working experiences of the military staff also significant to be considered in order to have an intuitive sense during the decision making. The experts will be able to create and develop their own mental models by exploring and discovering the previous mission. Refer Figure 4.0.

iii. The necessity of system/tool to support the process of decision making. As many (17%) stated that very critical and (33.3%) critical to implement the Multi-touch Table in decision making. Refer Figure 5.0.

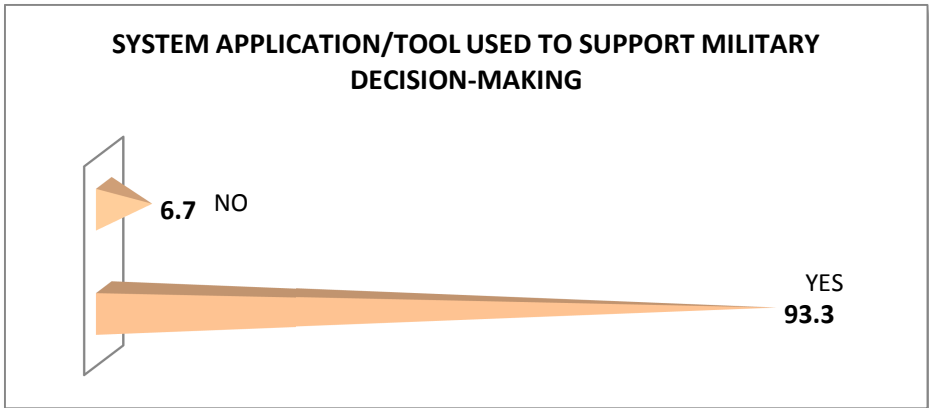


Fig. 3. The percentage of system/tool used in supporting the decision making

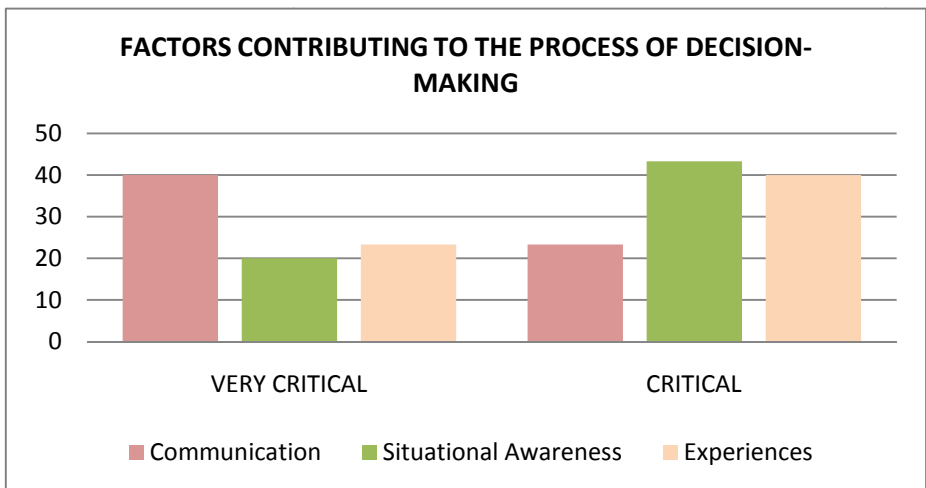


Fig. 4. Factors that influence the process of decision-making in the Army

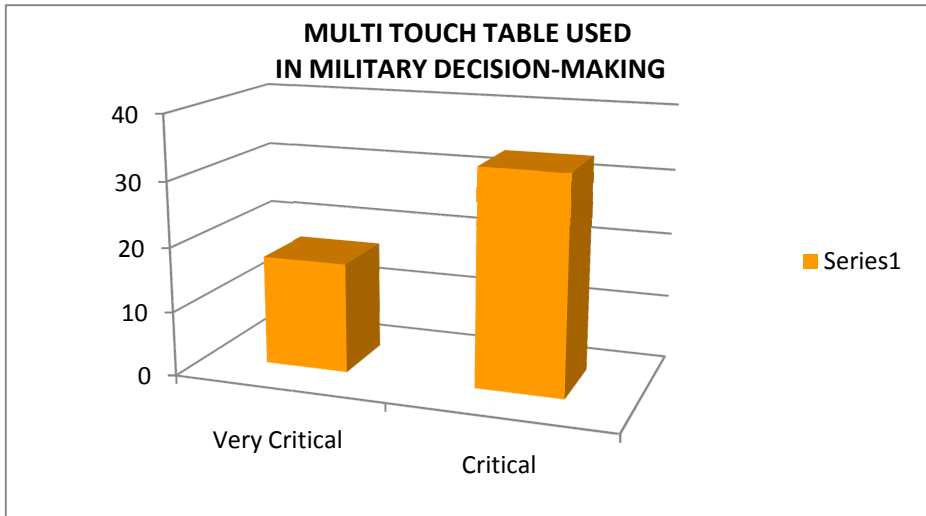


Fig. 5. Practicality of Multi-touch Table in the Military Decision Making

4 Conclusion

Decision making in such high stress and timely environments as a military operations center requires the concept of visualization in order to help the Commanding Officers (COs) and staff to visualize the scenarios, solutions and impact of the decision/COA. The decision making process that involved situations such as operating in an environment of volatility, uncertainty, chaos and ambiguity complicates the communication and interaction amongst militaries of higher and lower level. By investigating the effects of communication and interaction may give advantages to structure and develop the content using Multi-touch Table. Moreover, these Multi-touch technologies will allow multiple users to interact simultaneously in the most natural and intuitive method by only using their bare hands. Data and information also may be presented more efficient and effective ways due to this medium that can provide the nature of the touch-based interaction in order to quickly grasp complex controls using features provided such as zoom-in, zoom-out, drag and drop function.

As a conclusion, the adaptation of all the elements in Critical Success Factors (CSFs) that been identified through preliminary study in environment of new display technologies, example Multi-touch Table will allows for more intimate communication and interaction between Commanding Officer (CO) and others military staffs. This may leveraging the knowledge of multiple users in military decision making doctrine. This will also allows them to work together more efficiently and take full advantages of these new technology's abilities in the collaborative setting.

References

1. Burns, K.: Mental Models And Normal Errors In Naturalistic Decision Making. In: 5th Conference on Naturalistic Decision Making, Tammsvik, Sweden, May 26-28 (2000)
2. Daley, A.C., Harris, D.: Training decision making using serious games. Human Factors Integration Defense Technology Centre (2007)
3. Scott, W., Keith, K., et al.: Enabling Battlefield Visualization: An Agent Based Information Management Approach. 10th International Command and Control Research and Technology Symposium the Future of C2 (2010)
4. George, L., et al.: Synchronizing knowledge in Military Decision Making: A Research Approach for exploring the effects of organizational cultures. Information Systems IX(2) (2008)
5. Hancock, et al.: Informing the Design of Direct-Touch Tabletops. IEEE Computer Graphics and Applications 26(5) (2006)
6. Cummings, et al.: Vispol: An Interactive Tabletop Graph Visualization for the Police. In: Proceedings of ACM CHI 2011 Conference on Human Factors in Computing Systems (2011)
7. Endsley, M.R., Garland, D.J. (eds.): Theoretical Underpinnings of Situation Awareness: A Critical Review. Situational Awareness Analysis & Measurement. Lawrence Erlbaum Associates, Mahwah (2000)

Design of a Visual Query Language for Geographic Information System on a Touch Screen

Siju Wu¹, Samir Otmane¹, Guillaume Moreau², and Myriam Servières²

¹ IBISC, Université d'Evry Val d'Essonne, France

{siju.wu, samir.otmane}@ibisc.univ-evry.fr

² LUNAM Université, Ecole Centrale de Nantes – CERMA, France

{guillaume.moreau, myriam.servieres}@ec-nantes.fr

Abstract. This paper presents two spatial query methods for a Geographic Information System (GIS) that runs on a touch screen. On conventional GIS interfaces SQL is used to construct spatial queries. However keyboard typing proves to be inefficient on touch screens. Furthermore, SQL is not an easy-learning language, especially for novices to GIS. To simplify query construction, firstly we have designed a map interaction based query method (MIBQM). This method allows users to make simple queries by selecting necessary layers, features and query operators directly on the interface. To allow users to construct complex queries, a sketch drawing based query method (SDBQM) is proposed. Spatial query concepts can be represented by sketches of some symbolic graphical objects. It is possible to add spatial conditions and non-spatial conditions to describe query concepts more precisely. An evaluation has been made to compare SQL and MIBQM. We have found that for simple queries, MIBQM takes less time and proves to be more user-friendly.

Keywords: GIS, Touchable Interface, Visual Query Language, Spatial Query.

1 Introduction

Geographic Information System (GIS) is a system which is able to capture, store, analyze, manage and present the geographic referenced data [1]. Because of its powerful capability of data processing, now it has become an inevitable tool in many domains. In recent years touch screen technology has developed rapidly and has been widely used on Smartphones and Tablet PCs. Since the touch screen is apt to provide natural user experience, nowadays some GIS interfaces are redefined to adapt to gesture interaction. However, only some fundamental functions can be accomplished on these interfaces. There is no satisfying solution for spatial query construction. On conventional GIS interfaces such as ArcGIS [2] or OrbisGIS [3], SQL is used to make spatial queries. Unfortunately typing on a keyboard still remains inefficient on the touch screen. Besides this, users may find it difficult to translate query concepts into SQL statements in an intuitive way. A mature query construction method is still to be proposed.

In this paper we propose two spatial query methods which are apt to be used on touch screens. The first method is named map interaction based query method (MIBQM). This method allows users to construct simple queries by selecting necessary layers, features and query operators. If a specific feature is concerned in a query, it can be selected directly on the map. The second method is the sketch drawing based query method (SDBQM). By making use of a new visual query language, users can draw sketches to formulate queries. Different from MIBQM, spatial conditions and non-spatial conditions can be added to describe query concepts more precisely. These query methods are designed for specialists to improve their working efficiency. By applying these methods, query functions of GIS are also available for novices.

The remainder of this paper is organized as follows. In Sect. 2, we review related work about the design of spatial query methods. Then in Sect. 3 and Sect. 4 we present how to make queries by using our methods and we provide a brief introduction of the touchable interface in Sect. 5. In Sect. 6 we present a usability study which evaluates MIBQM by comparing it with SQL. Finally, we draw some conclusions and give a discussion about future research.

2 Related Works

On conventional interfaces, SQL is widely used as a tool to search spatial information. To give users the capability of dealing with spatial attributes of layers, different database systems have provided their own series of spatial functions. These functions can be used to construct SQL query statements. Although these abundant functions are powerful, some users may find it hard to remember all their names and application methods.

Besides SQL, there also exist other spatial query methods. One type of methods allows users to draw a flowchart to make a query. Users place flowchart elements, which can be a layer or a query operator, in a design space and then connect them in a manner consistent with the interface to construct a model [4] [5] [6]. A flowchart can express query concepts precisely, but it lacks visual explanation of the query concept. The flowchart may be complex to understand when the query is complex. To make GIS more accessible for non-trained users, visual query languages have been proposed. For many database management systems, visual query languages are often designed to improve the effectiveness of human-computer communication [7]. Some visual query languages allow users to use some predefined icons to compose spatial queries [8] [9]. Each icon thus represents a specific layer. Spatial relations between two layers can be represented by a relationship operator and query operators are used to indicate what kinds of data to search. One disadvantage of these methods is that once a new layer is added in the database, a new icon should be defined. Some other visual query languages offer more liberty to draw sketches. To represent a layer, users can draw a symbolic object and add a name to it [10] [11]. A spatial condition can be represented by the spatial relationship between two symbolic objects. The use of graphical representations offers an intuitive and incremental view of spatial queries, but it sometimes causes different interpretations of the same query. Users' query

concepts may be misunderstood by the system. Though most of the methods mentioned above are not designed for touch screens, some design concepts can still be borrowed. After analyzing these methods, we have proposed our spatial query methods.

3 Map Interaction Based Query Method

When making a spatial query, three kinds of information should be offered: what kind of data to search, which layers the data comes from and what conditions should be respected. The first two kinds of information are mandatory while the third one is optional. Inspired by a query method which allows users to find articles related to a certain subject directly through gesture interaction with the map [12], we have proposed MIBQM. Users can construct a spatial query by selecting necessary information which can be a layer, a feature or an operator on the interface and through combination of selected information, the system can automatically create a SQL statement.

In this method, two kinds of query operators are provided: unary operators (applicable to a single layer) and binary operators (that use two layers). Unary operators are used to get spatial attributes of a layer, such as the start points of rivers or the boundaries of provinces. Binary operators are for example called to calculate the topological relationships between features from two layers. To apply a unary operator, firstly a layer should be selected and the operator can be chosen. To apply a binary operator, users have to choose the first layer concerned in the query and the operator. After that the second layer should be selected. A query constructed in this way takes all the features of a layer into consideration. If only one feature is concerned in a query, after layer selection, this feature can be selected on the map by tap gesture. In this way, all the other features in the same layer will be ignored.

In this way, the map is no longer only used to consult numerical attributes, but can be exploited during the query construction. An example of calculation of the union of two province features is shown in Fig.1.

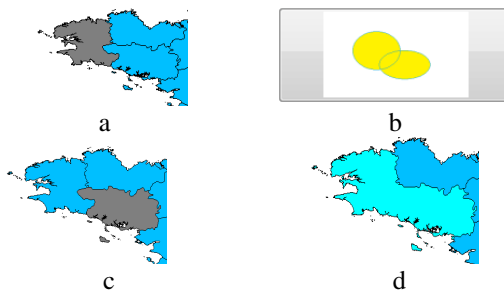


Fig. 1. Example of binary operator usage: a. select the first province b. select the union operator c. select the second province d. return the result

4 The Sketch Drawing Based Query Method

To make it possible to construct complex spatial queries, we have proposed another query method, which is named the sketch drawing based query method (SDBQM). This method allows users to use a new visual query language to represent query concepts by drawing sketches.

4.1 Symbolic Graphical Objects

For most people, it may be natural to describe their abstract concepts by drawing pictures because the meaning of pictures may resemble those concepts better than words. In SDBQM, symbolic graphical objects (SGO) can be used to specify queries (Fig.2). Each layer can be represented by a SGO in the drawing area. The appearance of a SGO may be a point, a line or a polygon according to its layer. A SGO is defined as a 5-tuple

$$\text{SGO} = \{\text{LAYER}, \text{ALIAS}, \text{SHAPE}, \text{POSITION}, \text{PROPERTYSET}\}$$

- LAYER is the name of the layer which is represented by the SGO;
- ALIAS is the alias name used to identify the SGO;
- SHAPE is the geometry shape of the SGO;
- POSITION is the location of the SGO in the drawing area;
- PROPERTYSET includes all the attributes of the layer which is represented by the SGO.

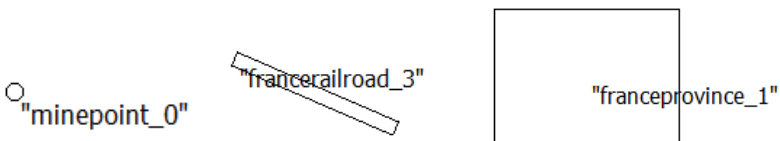


Fig. 2. SGO examples: Left: a point object; Middle: a line object; Right: a polygon object

4.2 Query Operators

After drawing the sketch, query operators can be added to make a query about a specific SGO. There are two kinds of operators that can be selected: unary-SGO operators and binary-SGO operators. Each operator corresponds to a specific spatial query. After using a selection envelope to cover the SGO involved in the query, an operator menu can be called.

Unary-SGO Operators

Unary-SGO operators (UOs) are used to called unary-parameter spatial functions (UFs), which are used to get attributes of a layer, such as buffers or boundaries. If a SGO is covered by the envelope, available UOs are displayed in the operator menu (Fig.3). After selecting one operator, the corresponding function will be called. The

geometry shape of the layer of the SGO is taken as parameter in the function. A new SGO is drawn to represent the result of the operator. By observing the appearance of the new SGO, users will know which operator has been selected.

Binary-SGO Operators

Similarly Binary-SGO operators (BOs) are used to call binary-parameter spatial functions (BFs) which can be used to calculate the geometric relationships between two layers. To display BOs in the operator menu, two SGOs should be covered by the envelope. Once a BO is selected, the corresponding duo-parameter spatial function will be called. The geometry shapes of the two layers of the SGOs are taken as parameters. And a new SGO is drawn to represent the result of the operator.

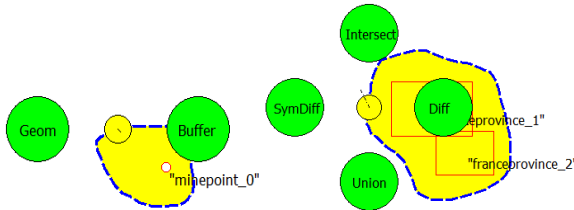


Fig. 3. SGO operators' example: Left: menu of UOs; Right: menu of BOs

4.3 Additional Conditions

To give users the ability to express complex query concepts, it is possible to add additional conditions. Additional conditions are classified into two groups: spatial conditions and non-spatial conditions. Spatial conditions are used to describe topological relationships that features should satisfy, while non-spatial conditions give constraints about numerical attributes of features.

Spatial Conditions

A spatial condition can be expressed by the topological relationship between two SGOs in the drawing area. Five relationships can be identified: DISJOINT, TOUCH, INTERSECT, EQUAL and COVER. Drawing SGOs in specific relationships to express spatial conditions affords an intuitive and comprehensible view of spatial queries. However, if each pair of relationship is translated into spatial conditions, ambiguity may appear. To avoid those ambiguities, a topological table is used. In the topological table each SGO is represented as a circle with its name. If the topological relation between two SGOs should be translated into a spatial condition, a connection line can be drawn between them. To remove a spatial condition, a cut line can be drawn to cancel the connection. In this way the meaning of the sketches can be explained more precisely. The example in Fig 4 has added two spatial conditions: franceprovince_0 has an intersection with franceprovince_1 and francerailroad_2 crosses franceprovince_0. The relationship between franceprovince_1 and fracerailroad_2 is ignored.

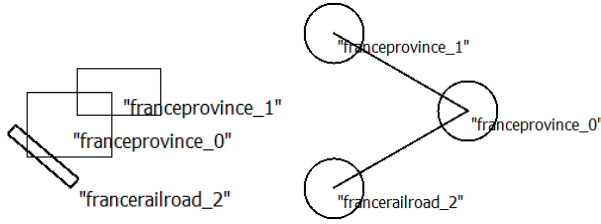


Fig. 4. Left: A sketch of three SGOs; Right: the corresponding topological table

Non-spatial Conditions

In SDBQM, we have designed non-spatial condition objects (NCOs) and condition connector objects (CCOs) to add non-spatial conditions. A NCO is defined as 4-tuplet

$$NCO = \{LAYER, ATTRIBUTE, OPERATOR, VALUE\}$$

- LAYER is the name of the layer concerned in the condition;
- ATTRIBUTE is the attribute selected in LAYER;
- OPERATOR is used to decide which kind of constraint is set about ATTRIBUTE;
- VALUE is used to compare with ATTRIBUTE.

If more than one NCO is added, CCOs can be used to organize different NCOs in a nested structure. A CCO is defined as 3-tuplet

$$CCO = \{TYPE, NODE_{FIRST}, NODE_{SECOND}\}$$

- TYPE is the connection type. There are two types of CCO: AND or OR;
- NODE_{FIRST} is the first object of the connection. It can be a NCO or a CCO;
- NODE_{SECOND} is the second object of the connection. It can be a NCO or a CCO.

Three NCOs are organized in a nested structure in Fig.5 and two CCOs are used to connect them.

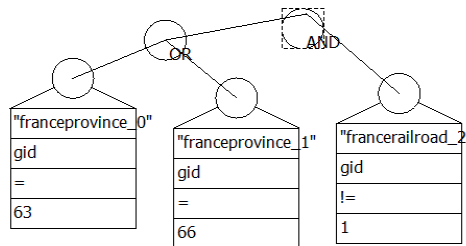


Fig. 5. Nested structure using conditions

4.4 Composed Query

After a query is executed, a new layer is generated and the query result is saved in this layer. Since the geometry types of features in the result layer may be different, it is

impossible to use a classic geometry shape to represent the SGO of the new layer. Our solution of this problem is to use the shape of the result SGO which is generated in the last query. In this way, the appearance of the result SGO may be comprehensible and meaningful to recall what query has been done in the last time.

5 Touchable Interface

To realize the two query methods proposed in this paper, we have developed a touchable GIS interface. The interface consists of two parts: the basic interface (BI) and the sketch query interface (SQI). On BI, fundamental functions such as map manipulation and layer manipulation can be accomplished. On the right side of BI a toolbox is set. Three tags are used to display functions of different types. In the third tag spatial operators of MIBQM are provided. To apply SDBQM, SQI should be used. On SQI, there is a drawing area in which sketches can be drawn. Two condition-areas are set. One is used to set the topological table, and another is used to add NCOs.

On the touchable interface, users can apply various gestures for different tasks. In BI, gestures are used to manipulate the map. A drag gesture with one finger can translate the map and a pinch gesture with two fingers can scale the map. To rotate the map, users have to first press two fingers on the map to set their center as the rotation center, and then release one finger. The map can be rotated around the center by the remaining finger. In this way rotation and zoom manipulation can be separated. We have also designed a three-finger pan gesture to tilt the map in the horizontal or vertical direction. In BI, the data table can be called or hidden by a four-finger vertical gesture. In SQI, all the tasks such as drawing of SGOs, selection of operators and addition of conditions can be accomplished by gestures. To draw a SGO, users firstly have to select a layer in the layer list. If it is a point-shape layer, the SGO can be drawn by a tap gesture in the drawing area. Else the SGO of a line-shape or polygon-shape layer can be drawn by a drag gesture. After sketches are drawn, the drawing of a selection envelope can be started by pressing one finger in the drawing area. If the envelope crosses its start point and there are one or two SGOs covered by the envelope, the operator menu will be displayed. In the non-spatial condition area, a three-finger tap gesture can add an empty NCO. To switch from BI to SQI, users can drag five fingers from left to right in the map browsing window. Similarly a five-finger drag from right to left will hide SQI and recall BI.

6 Evaluation

We have performed an evaluation to examine the functionality of MIBQM. SDBQM will be evaluated in the future work. We asked 10 testers (8 male and 2 female) to accomplish 3 spatial query tasks by using SQL and MIBQM respectively. The average age of these testers is 26. All the testers are novices to GIS and only 2 of them are familiar with SQL. All three tasks are listed as follows.

1. Get the geometry shapes of all the features in the railway layer of France
2. Calculate the buffers of all the features in the mine field layer of France
3. Calculate the intersection of two provinces of France

Because the SQL console is not implemented in the touchable interface, so testers constructed SQL statements on OrbisGIS, which is an open source GIS. All the testers are divided into two groups. People in the first group tested SQL first and then MIBQM, while people in the second group did the inverse order. Before the evaluation, they were told how to accomplish these tasks and gave them some time to attempt the solution of each task. To compare the learning rate of these two query methods, each task has been repeated for three times. During the test, the performance time and the number of errors are measured. After the evaluation, each tester has answered a questionnaire to collect their comments.

6.1 Objective Evaluation

From Table 1 we find out that MIBQM takes less time to fulfill all the tasks. When using SQL, testers spent most of the time in typing statements and consulting attribute values. However, MIBQM only necessitates a few of gestures. For SQL, the mean time for each task is 20s, 31s and 2:04min, and the standard deviations are 10s, 13s and 1:06min. For MIBQM, the mean time for each task is 7s, 21s and 20s, and the standard deviations are 5s, 10s and 11s. Performance improvements are 32%, 32% and 47% for SQL. For MIBQM improvements are 58%, 4% and 38%.

The table in Fig.6 shows that testers have made fewer errors for task1 and task2 when using MIBQM. Freshmen of SQL may find it hard to construct statements without grammar errors or false input of layer names and attribute values. However, for task3 MIBQM leads to more errors. Before feature selection, some users forgot to select the layer, so they found it impossible to select a feature. Each tap without a successful selection is considered as an error and it is why there are more errors for MIBQM. Some false selections have been made because some users wanted to select a feature when the size of map is not large enough. Some users confused the order of feature selection and the order of operator selection, so wrong results were obtained. Most of the errors with MIBQM are found and corrected by testers before running the query, while errors with SQL are found by the system after testers ran the query.

Table 1. Average performance time

Time Task	Test1SQL	Test2 SQL	Test3 SQL	Test1 MIBQM	Test2 MIBQM	Test3 MIBQM
Task1	0:25	0:20	0:17	0:12	0:06	0:05
Task2	0:37	0:33	0:25	0:24	0:18	0:23
Task3	2:55	1:45	1:33	0:26	0:18	0:16

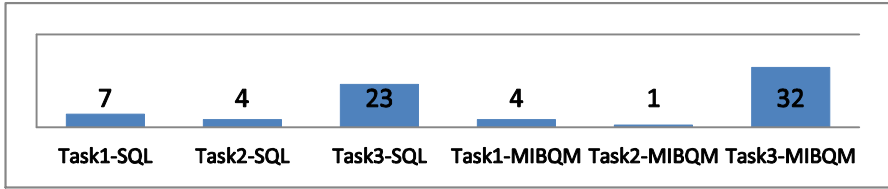


Fig. 6. Testers' number of errors

6.2 Subjective Evaluation

In terms of usability, we found out that 80% testers think MIBQM is easy to use and 80% testers think features are easy to be selected. And in terms of satisfaction, all the testers agree that it is easier to add a condition by selecting a feature on the map than by inputting the primary key of the feature in the SQL statement. 80% of testers think that MIBQM performs better in the aspect of leading to fewer errors. All the testers prefer MIBQM to SQL.

Some problems of MIBQM are found during the evaluation. More user guidelines are expected to make the query procedures clearer, so that users will not be confused by the procedure orders. The feature selection method should be improved to avoid invalid selection. For the calculation of topological relationship, one tester thought it more reasonable to make operator selection before layer selection and suggested to use multi-features selection. So in the future work we will focus on these problems to improve the user experience of MIBQM.

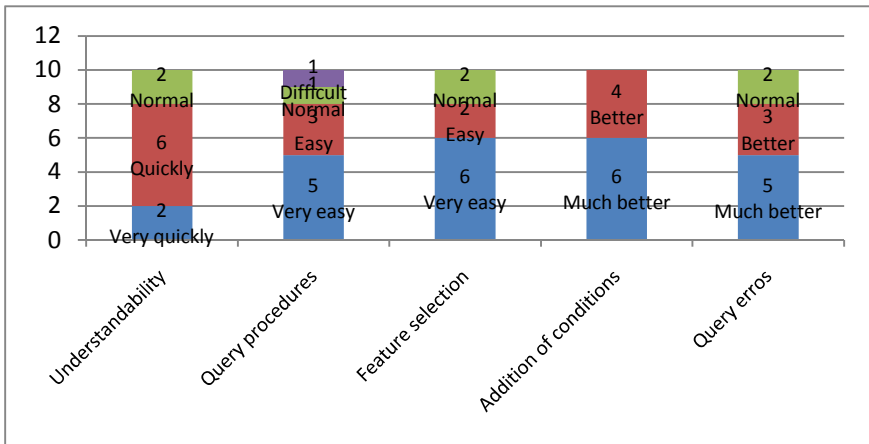


Fig. 7. Assessment of MIBQM

7 Conclusion

In this paper we have presented two spatial query methods which are realized on a touchable GIS interface. MIBQM allows users to construct spatial queries by selecting

layers and query operators. If a query is made about a specific feature, users can directly select the feature on the map. All the other features in the same layer will be ignored. To make it possible to describe query concepts more precisely, additional conditions should be offered. SDBQM permits spatial query construction by drawing sketches of SGOs. If two SGOs are connected in the spatial relation table, a spatial condition will be added according to their topological relationship. Users can also draw NCOs to add non-spatial conditions. NCOs can be organized in a nested structure by adding CCOs. If a query is too complex, it can be separated to several simple queries. Layers generated in queries can also be represented by special SGOs to be reused. We have made an evaluation between SQL and MIBQM and we found that for novices to GIS, MIBQM seems easier to understand and apply. For simple spatial query tasks, using MIBQM takes less time and leads to fewer errors. In the future, we will compare the SQL and SDQM through constructing complex spatial queries. We also hope to improve MIBQM so that additional conditions can be added.

References

1. Denègre, J., Salgé, F.: Les systèmes d'information géographiques. In: Presses Universitaires de France, 2nd edn. (2004)
2. ArcGIS, <http://www.esri.com/software/arcgis/arcgis-for-desktop>
3. OrbisGIS, <http://www.orbisgis.org/>
4. Kirby, K., Paner, M.: Graphic Map Algebra. In: Brassel, K., Kishimoto, H. (eds.) Proceedings of the 4th International Symposium on Spatial Data Handling, Zurich, Switzerland, pp. 413–422 (1990)
5. Lanter, D., Essinger, R.: User-Centered Graphical User Interface Design for GIS. In: Technical Report 91-6, Santa Barbara, CA: National Center for Geographic Information and Analysis (1991)
6. ERDAS, Model Maker Tour Guide. Atlanta, GA: ERDAS, Inc. (1993)
7. Catarci, T., Costabile, M.F., Levialdi, S., Batini, C.: Visual query systems for databases: a survey. *Journal of Visual Languages and Computing* 8, 215–260 (1997)
8. Calcinelli, D., Mainguenaud, M.: Cigales, a Visual Query Language for a Geographical Information System: the User Interface. *Journal of Visual Languages and Computing* 5(2), 113–132 (1994)
9. Sebillio, M., Tortora, G., Vitiello, G.: The Metaphor GIS Query Language. *Journal of Visual Languages and Computing* 11, 439–454 (2000)
10. Ferri, F., Rafanelli, M.: Resolution of Ambiguities in Query Interpretation for Geographical Pictorial Query Languages. *Journal of Computing and Information Technology* 12(2), 119–126 (2004)
11. Egenhofer, M.J.: Query Processing in Spatial-Query-by-Sketch. *Journal of Visual Languages and Computing* 8(4), 403–424 (1997)
12. Schoning, J., Raubal, M., Marsh, M., Hecht, B., Kruger, A., Rhos, M.: Improving Interaction with Virtual Globes through Spatial Thinking: Helping users Ask ‘Why?’. In: Proceedings of the 13th Annual ACM Conference on Intelligent User Interfaces. ACM, USA (2008)

Target Orientation Effects on Movement Time in Rapid Aiming Tasks

Yugang Zhang, Bifeng Song, and Wensheng Min

School of Aeronautics, Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China
zhang_yu9999@163.com

Abstract. An attempt was made to investigate the effect of the target orientation on pointing performance. An experiment was accomplished in which 10 subjects performed three-dimensional aiming tasks under the manipulation of target orientation, distance to target and direction to target. Results show that target orientation affects the duration of three-dimensional movements significantly. As a result, the conventional movement model did not satisfactorily explain the variance in the movement times produced. The conventional model was employed by incorporating an oriented parameter into the model. The modified model was shown to better fit the data than the conventional model, in terms of r^2 between the measured movement time and the value predicted by model fit.

Keywords: Human movement, Pointing performance, Fitts' Law, Index of difficulty, Target orientation.

1 Introduction

Modern human computer interfaces (for example, the touch screen, and the three-dimensional display) are becoming much more popular so it is important to determine whether current movement model can provide useful predictions for these interfaces as well [1,13]. Since the famous speed-accuracy model of human movement, Fitts' law, was developed, many researchers have verified it over a wide range of conditions [11] and applied in primarily two ways, as a predictive model, and as a means of the comparison and evaluation of pointing devices [7,12].

Fitts' law as originally is a one-dimensional model of human movement. It predicts the movement time MT to select a target of width W and distance (or amplitude) A from the starting point. MacKenzie and Buxton [8] extended it to two-dimensional tasks. They applied target amplitude A , target width W , target height H , and approach angle θ_A to their two-dimensional model. Then Accot and Zhai [2] further investigated bivariate pointing based on Fitts' law model. They focused on the effect of target shape (target width and height ratio) on pointing performance. Moreover, Murata and Iwase [10] extended Fitts' law to three-dimensional pointing tasks and incorporated a directional parameter into the model. In contrast, Grossman and Balakrishnan [5] proposed a new model that describes pointing at trivariate targets. In addition,

researchers had studied many other factors, such as trade off of speed and accuracy [4], target scale [6]. However, previous records have failed to consider target orientation. Thus, the results are controversial when apply typical model to three-dimensional pointing tasks. Uncertainties still exist.

The target orientation means, in real world, the target object (e.g., a button, switch, even “iconic” menu) is lying on a surface and cannot exist solely. The plane of target and the plane of starting point present an angle, which we define as the “target orientation” (see Fig. 1 for an illustration).

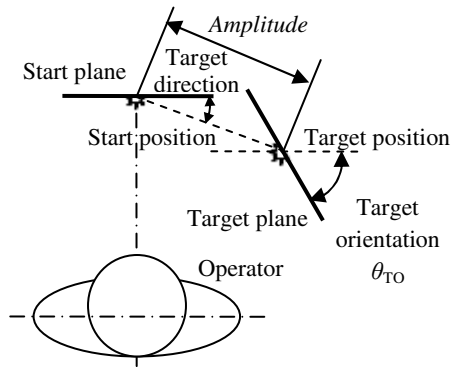


Fig. 1. Sample operation condition

The present study concerns the effect of the target orientation on pointing performance. The control of the forces over the amplitude becomes far complicated with a present of the target orientation. The task is a three-dimensional pointing task and exacter muscular force is required, leading to more variable movement trajectories and, hence, more variable pointing times [14]. Based on these insights into aiming movements, we argued that the pointing movements studied in the present experiment would be sensitive to the effects of target orientation. Fitts' law would be out of action.

In this paper, we studied to examine this hypothesis and to model operator performance in the most fundamental interaction task – pointing – in an experiment where the target orientation varied. To anticipate, we realized this goal as follows.

First, we used the conventional Fitts' model to predict movement time data collected in a pointing task under the manipulation of target orientation, distance to target and direction to target. The fit was suboptimal due to the variance present in the data and the dependency of movement time on target orientation. Second, based on these results, a modified three-dimensional model of Fitts' law was proposed, which was shown to describe the data better than the conventional Fitts' model. Third, we investigated the effects of the factors identified. Finally, we concluded by discussing implications for user interface design.

2 Methods

We will provide in this section experiment method which are necessary for the understanding of subsequent results.

2.1 Subjects

Ten Northwestern Polytechnical University healthy male postgraduates (24–26 years of age) participated in the experiment. The subjects were all right-handed and inexperienced with regard to the purpose of the experiment.

2.2 Apparatus

The experiment was conducted on Lenovo[®] PC equipped with two 21.5" touch screen LCD monitor (47.8cm×27cm visual area, 1024×768 pixels, 96 dpi resolution). The two monitor was placed vertically on the same shelf (see Fig. 2). One of them showed the starting point and another showed the target point.



Fig. 2. Physical setup for a pointing task

2.3 Task

The task's paradigm is discrete task. The starting point was placed on the median sagittal plane of the subject, which was about 60 cm away in front of the subject. The subject was required to place his right index finger at the starting point before the experimenter gave him the signal to start the movement. The subject's task was to point with the right index finger to the target specified by the experimenter.

A two-dimensional circle was used as the target to equalize the distance between the starting point and the target for each orientation condition with equal values of target size and distance. In this study, three-dimensional pointing means that the

movement of the pointer (tip of the index finger) is performed in a three-dimensional space and measured along three axes.

2.4 Design and Procedure

A within-subject factorial design with repeated measures was used. The independent variables were the distance to target (three levels: 150, 300, 450mm), target orientation (four levels: 0°, 30°, 60°, 90°) and direction to target (two levels: 0°, 30°). The target width is 18mm. Dependent variables were movement time (*MT*). The error rate was controlled at about 4%. There were 24 different combinations in total.

The experiment included two sessions: a practice session, to allow participants to get used to the task and conditions, and a data-collection session, wherein participants tested the 24 different combinations in certain order. Within each condition, participants performed 20 trials.

After having signed an information consent statement, each subject was tested. The subjects were instructed to carry out the task as accurately and as quickly as possible. The time when the fingertip began to leave from the screen of the starting point was used as a criterion for movement onset. The criterion indicating the end of the movement (trial) was the time when the tip of the index finger reached the screen of the target point. The movement (pointing) time was obtained using developed test software. If the coordinate was within or on the target circle, then the trial was regarded as successful. All other cases were designated as error trials.

3 Results

Outliers were got rid of based on mean movement time and accuracy – defined as distance between the click point and the target center. Any data further than 3 standard deviations away from its condition's mean (by *MT*) was removed. 0.9% of the data were removed as outliers.

3.1 Movement Time Analysis

Figure 3 shows how mean movement time changes as a function of the set of amplitude tested in our experiment, while all other factors are balanced. In the figure, “0° Motion” means the direction to target is 0° and “30° Motion” means 30°. Immediately noticeable is that the mean movement time (averaged over all orientation conditions and all subjects) generally increases with the increasing distance to target. The movement time of 30° target direction is shorter than that of 0° target direction. Analysis of variance showed that the independent variable *A* ($F_{2, 18} = 1862, p < .01$) has a significant effect on *MT*. This is consistent with the findings of previous studies [3].

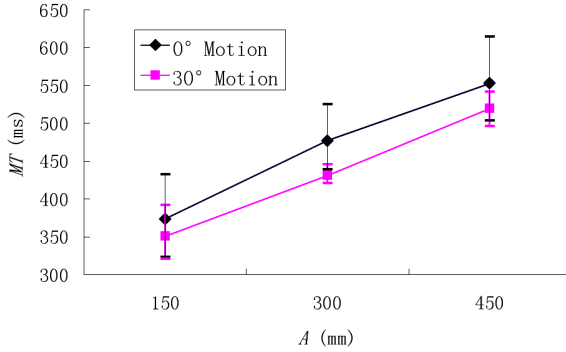


Fig. 3. Influence of target relative position on MT

3.2 Fitting the Data to the Conventional Fitts' Model

First, the data were modeled using the following conventional Fitts' model [9]:

$$MT = a + b \log_2(A/W + 1.0) \tag{1}$$

Where MT represents the time to move the right index finger from the starting point to the target, and A and W are the distance from the starting point to the target and the size (diameter) of the target, respectively. The term $\log_2(A/W + 1.0)$ is the index of difficulty carrying the unit of bits. Finally, the parameters a and b are empirical constants to be determined through linear regression.

The mean MT was calculated for each index of difficulty, pooled over all orientation conditions and subjects. The r^2 of the linear regression between mean MT and index of difficulty was 0.789 (see Fig. 4). On the basis of this relatively poor fit it can be concluded that there is still substantial room for improving upon the conventional Fitts' model when it comes to the description of three-dimensional movements.

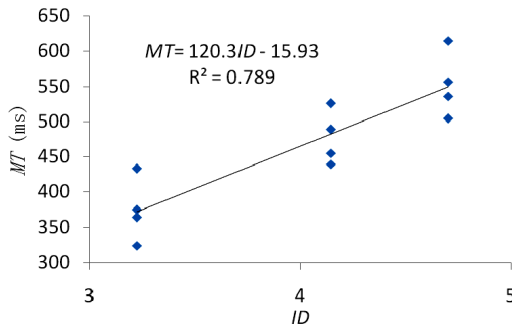


Fig. 4. Relationship between index of difficulty (ID) and movement time

3.3 Modifying Fitts' Law to a Three-Dimensional Pointing Task

To find a meaningful extension of the conventional Fitts' model to three-dimensional pointing movements, we went on to examine the expected relationship between movement time and target orientation.

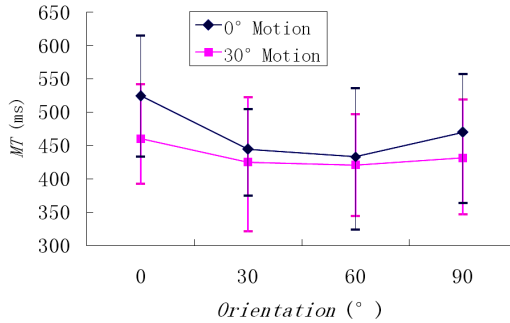


Fig. 5. Influence of target orientation on MT

Figure 5 shows how the mean movement time MT (averaged over all amplitude conditions and all subjects) varied across the four levels of target orientation. The tendency like a “bowl” shape. The subjective impression from this figure that MT depended on orientation was confirmed in a one-way ANOVA, which revealed a significant main effect of θ_{TO} ($F_{3,27}=167.198, p < .01$). To identify the source of this statistically significant effect, a multiple comparison post hoc test was performed (i.e., Least Significant Difference test) using a conservative significance level of $p < .01$. On this test, all the comparisons were significant. These results indicate the presence of a systematic relationship between movement time and the target orientation. These imply that a model considering orientation will lead to a better performance model than the conventional Fitts' model. Therefore, we judged that the target orientation could be taken into account by incorporating θ_{TO} into the ID in Eq. (1). Based on the discussion, the ID was revised using the following formula:

$$ID_{TO} = \log_2(A/W + 1.0) - c \sin(2 \times \theta_{TO}) \tag{2}$$

Where c is an arbitrary constant to be determined through linear regression.

For several values of c , the relationship between ID_{TO} and movement time MT was established by means of linear regression. However, the target orientation seems to be an important factor in performance modeling, especially the modeling of three-dimensional pointing tasks. Figure 6 shows the r^2 for the data in Figure 4 as a function of c . The highest r^2 (0.913) was found for $c=0.5$. The fit to the experimental data was improved by using the index of difficulty ID_{TO} , which incorporates the effect of orientation on MT , and by using the value of c producing the highest r^2 (Fig. 7). The optimal values of c differed for the experimental data of the individual subjects.

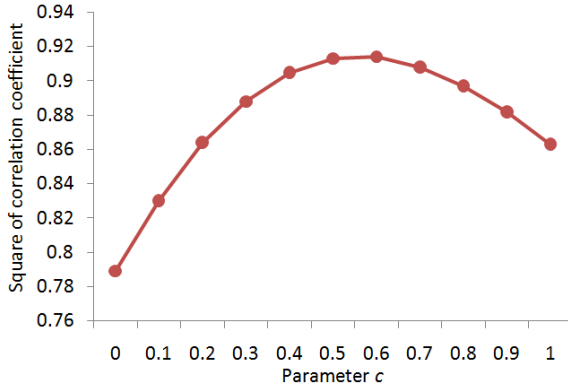


Fig. 6. Squared correlation coefficients as a function parameter c for fitting the data

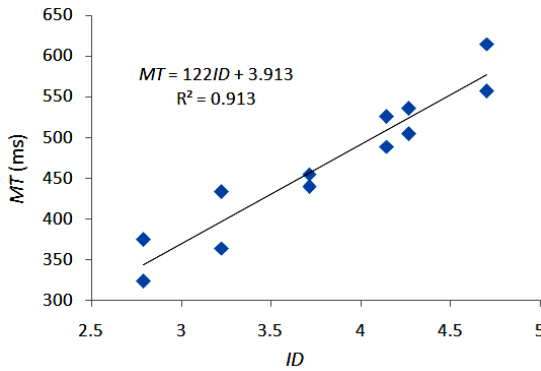


Fig. 7. Movement time as a function of the ID in the modified Fitts' model ($c=0.5$)

The fit to the obtained movement times was better when extended three-dimensional modeling was applied. A one-way (modeling method: conventional Fitts' law vs. extended three-dimensional modeling) ANOVA as used for the r^2 values showed that the difference was significant ($F_{1,9}=53.685, p < .01$). Collectively, these results clearly indicate that the modified model of Fitts' law better predicts the duration of three-dimensional (pointing) movements than the conventional Fitts' model.

4 Conclusions

The result in Figure 4 and Figure 7 illustrate an issue with traditional applications of Fitts' law to Three-dimensional Pointing Tasks. In the present experiment, movement time was affected significantly by target orientation (cf. Fig. 5). A tendency of “bowl” shape was found. It is not the shortest movement time when target object is lying on the same surface (target orientation is 0°) of starting point.

Thus, the conventional Fitts' model cannot adequately explain the variance in movement time in a real-world three-dimensional pointing task, as it does not take the target orientation into account (cf. Fig. 4). We can come to a conclusion that the interpretation of target orientation and the formulation used in the calculation of a task's index of difficulty play a critical role in the accuracy of the model.

Our study investigated how target orientation affect selection performance, and the results can provide us with significant guidelines on the layout design of car, airplane, and manipulator device, i.e. how items such as widgets, menus, and other objects should be sized and positioned in 3D layout.

In the present research, four levels of target orientation were employed. In future research, more graded levels of orientation will be used in order to confirm the reproducibility of our results. In other words, the relationship between the movement time and the target orientation must be confirmed using an experimental paradigm with more than four levels of target orientation. Furthermore, the subjects in the present study were all right-handed. Future research may investigate modeling that involves right-handed subjects performing pointing tasks with their left hands and left-handed subjects participating in the experiments. Finally, in the present experiment the pointing was conducted, for the sake of simplicity, by using a two-dimensional target in a three-dimensional space. In future research, pointing movements to three-dimensional targets should be examined, requiring a method for defining the target size for a sphere and bringing in depth to the task.

Acknowledgements. This work is supported by National Basic Research Program of China (973 Program), 2010 CB734101 and 2012 undergraduate final paper support program of Northwestern Polytechnical University.

References

1. Abdulin, E.: Using the Keystroke-Level Model for Designing User Interface on Middle-Sized Touch Screens. In: Proceedings of CHI 2011 Conference on Text Entry & Typing, Vancouver, BC, Canada, May 7–12, pp. 673–686 (2011)
2. Accot, J., Zhai, S.: Refining Fitts' Law Models for Bivariate Pointing. In: Proceedings of CHI 2003 Conference on Human Factors in Computing Systems, Ft. Lauderdale, Florida, USA, April 5–10, pp. 193–200 (2003)
3. Boritz, J., Booth, K.S., Cowan, W.B.: Fitts' law studies of directional mouse movement. In: Proceedings of Graphics Interface, Toronto, pp. 216–223 (1991)
4. Dean, M., Wu, S.-W., Maloney, L.T.: Trading off speed and accuracy in rapid, goal-directed movements. *Journal of Vision* 7(5),10, 1–12 (2007)
5. Grossman, T., Balakrishnan, R.: Pointing at Trivariate Targets in 3D Environments. In: Proceedings of CHI 2004 Conference on Human Factors in Computing Systems, Vienna, Austria, April 24–29, pp. 447–454 (2004)
6. Guiard, Y.: The Problem of Consistency in the Design of Fitts' Law Experiments: Consider either Target Distance and Width or Movement Form and Scale. In: Proceedings of the 27th International Conference on Human Factor in Computing Systems, Boston, Massachusetts, USA, April 4–9, pp. 1809–1818 (2009)

7. MacKenzie, I.S., Isokoski, P.: Fitts' Throughput and the Speed-Accuracy Tradeoff. In: Proceedings of CHI 2008 Conference on Human Factors in Computing Systems, Florence, Italy, April 5-10, pp. 1633-1636 (2008)
8. MacKenzie, I.S., Buxton, W.: Extending Fitts' Law to Two-dimensional Tasks. In: Proceedings of CHI 1992 Conference on Human Factors in Computing Systems, New York, USA, May 3-7, pp. 219-226 (1992)
9. MacKenzie, I.S.: A note on the information-theoretic basics for Fitts' law. *Journal of Motor Behavior* 21, 323-330 (1989)
10. Murata, A., Iwase, H.: Extending Fitts' law to a three-dimensional pointing task. *Human Movement Science* 20, 791-805 (2001)
11. Plamondon, R., Alimi, A.M.: Speed/accuracy trade-offs in target-directed movements. *Behavioural and Brain Sciences* 20, 279-349 (1997)
12. Soukoreff, R.W., MacKenzie, I.S.: Towards a standard for pointing device evaluation, perspectives on 27 years of Fitts' law research in HCI. *Int. J. Human-Computer Studies* 61, 751-789 (2004)
13. Teather, R.J., Stuerzlinger, W.: Target pointing in 3D user interfaces. In: Proceedings of the Graphics Interface Poster Session (GI 2010), Ottawa, Ontario, Canada, May 31-June 2, pp. 20-21 (2010)
14. VanGalen, G.P., DeJong, W.P.: Fitts' law as the outcome of a dynamic noise filtering model of motor control. *Human Movement Science* 14, 539-571 (1995)

Part IV
Haptic Interaction

Comparison of Enhanced Visual and Haptic Features in a Virtual Reality-Based Haptic Simulation

Michael Clamann, Wenqi Ma, and David B. Kaber

Edwards P. Fitts Department of Industrial and Systems Engineering,
North Carolina State University, Raleigh, NC, USA
{mpclamann, wma4, dbkaber}@ncsu.edu

Abstract. An experiment was conducted to compare the learning effects following motor skill training using three types of virtual reality simulations. Training and testing were presented using virtual reality (VR) and standardized forms of existing psychomotor tests, respectively. The VR training simulations included haptic, visual and a combination of haptic and visual assistance designed to accelerate training. A comparison of performance test results prior to and following training revealed conditions providing haptic assistance to yield lower scores related to fine motor skill training than the visual-only aiding condition. Similarly, training in the visual condition resulted in comparatively lower cognitive skill scores. The present investigation incorporating healthy subjects was designed as part of an ongoing research effort to provide insight on the design of VR simulations for rehabilitation of motor skills in patients with a history of mTBI.

Keywords: haptics, virtual reality, rehabilitation.

1 Introduction

In this research, we compared the effects of visual and haptic assistance for motor training using a virtual reality (VR)-based haptic simulation. Previous research on VR-based haptic simulation has demonstrated the efficacy of such tools for occupational therapy, including motor function rehabilitation [1-4]. The advantages for motor training include reducing trainer workload and training task costs by delivering any number of therapy sessions while maintaining accuracy and objectivity. Virtual reality can also enhance training by incorporating augmented controls and decision features during therapy sessions to aid user performance. Such features include precise corrective haptic control forces and enhanced visual aids that respond automatically to user actions. Similar enhancements would be difficult to implement in a physical system. However, questions remain on how to best implement these technologies and what specific VR design features might serve to accelerate motor learning beyond VR training tasks that merely replicate traditional training environments.

Previous research by our team identified combinations of augmented visual and haptic features that may provide therapeutic benefits over traditional VR systems [5]. The experiment design replicated a simplified occupational therapy regimen in which

a drawing and pattern assembly task represented occupational tasks that were anticipated to improve as a result of therapy. A VR reproduction of the block design (BD) subtest from the Wechsler Abbreviated Scale for Intelligence (WASI; [6]) was developed to be used to train subject motor skills in a course of simulated therapy. The BD subtest requires subjects to build replicas of patterns using blocks printed with simple patterns. Subjects are given a collection of nine red and white cubes with varying patterns on each side and are asked to replicate designs shown on a series of test cards. Scoring is based on speed and accuracy. In our study, healthy subjects were trained in the BD task using the non-dominant hand to simulate minor motor impairment. Training effects were measured by comparing drawing and pattern assembly task test scores obtained before (pre-test) and after (post-test) multiple BD training sessions. Subjects were assigned to one of three groups, including performance of the native BD task using standardized test materials (i.e., test cards and nine 1-inch cubes), use of a basic VR simulation of the task, or an augmented VR simulation with additional visual and haptic aiding. Results revealed a significant improvement in post-test performance over pre-test for the augmented VR training. In general, the study supported integrating haptic control in VR for psychomotor skill training. It also provided useful information for future haptic VR simulation design. However, because visual and haptic features were combined in the augmented condition, further investigation was needed to determine the extent to which these two forms of assistance contributed individually to psychomotor training.

Prior research comparing visual and haptic training modalities has produced mixed results. In one study [7], subjects were trained to replicate a 3-dimensional (D) trajectory using a haptic controller with 3 degrees of freedom (DOF). Subjects were initially required to trace the trajectory using visual, haptic, or a combination of visual and haptic conditions. Results showed that the haptic training alone was more effective with respect to timing as compared to visual training, but less effective with respect to absolute position and shape accuracy measures. Furthermore, the researchers also found that combining visual and haptic control during training did not provide additional learning beyond that of the visual-only condition.

Another research group [8] compared the effects of training visuomotor skills using combinations of haptic guidance and visual demonstration. Their methods were similar to those used in [7], but featured a simpler trajectory, additional training and additional test trials to more closely resemble an occupational therapy regimen. The researchers found that both visual-only and the combination of visual and haptic training allowed subjects to improve their ability to reproduce a novel trajectory. The results also showed that the combination of haptic and visual input did not significantly improve learning compared to the visual input alone, which supported the findings of [7]. Moreover, the researchers found that subjects receiving visual training performed marginally better than those receiving a combination of visual and haptic training. The authors speculated this occurred because visual feedback is more accurate than haptic; therefore, haptic feedback does not contribute to improved performance when both types of feedback are available at the same time.

These studies suggest that assistance that combines visual and haptic components may not be more effective than visual or haptic assistance alone. However, there are several differences between the procedures used in [7] and [7] and the present study. First, the task itself is quite different (i.e., tracking vs. pattern assembly). Second, in the prior studies the training and test tasks were the same. In the present study, in contrast, we implemented a training task designed to affect performance in a different test task, representing an occupation-related activity distinct from the training task.

The present study used the apparatus incorporated in the prior work, including a VR version of the Rey-Osterreith Complex Figure (ROCF; [9]) test to represent the occupational task and a VR-based haptic simulation of the BD (VR-BD) subtask from the WASI [6], representing a training task [10], [5].

The VR BD training task used a combination of haptic and visual aiding. Haptic assistance in the VR-BD included scheduled snap forces and rejection forces. The snap force was expected to assist users by prompting the correct movement when approaching a target block position at close range [11]. It was designed to reinforce correct placement and reduce the need for additional visual verification of block position. The rejection force, in contrast, was designed to reveal block placement errors. Both forms of haptic assistance were designed to assist users passively (i.e., without additional voluntary control of the haptic device).

Visual assistance providing passive positive feedback during correct block placement and corrective feedback during incorrect placements was also implemented. Subjects could also actively request assistance to “decompose” a design to reveal in individual block positions and orientations within a design. Use of this feature came at the cost of additional task time while activating the assistance. Precise details on the nature of the haptic and visual aiding are provided in the methods section below.

The current study investigated the influence of these augmented visual and haptic VR features, independently and in combination, on subjects learning. Based on the previous research, three different augmented VR conditions (i.e., haptic, visual or combined haptic and visual aiding) [12] were delivered to healthy subjects through a simplified occupational therapy regimen. This study also served as an additional step in the development of a proof of concept of the VR system to be used with patients with a history of minor traumatic brain injury (mTBI) for motor skill rehabilitation.

2 Methods

Twenty-four subjects between the ages of 18 and 44 were recruited for the study. All subjects were required to have 20/20 or corrected to normal vision and to exhibit right-hand dominance. Right-hand dominance was confirmed using the Edinburgh Handedness Inventory [13]. Subjects were required to complete all testing and training as part of the experiment using the left hand. This requirement was used to

simulate minor motor impairment and to disadvantage subject task performance in order to promote sensitivity to the training conditions.

The VR-BD task was presented on a PC integrated with a stereoscopic display using a NVIDIA® 3D Vision™ Kit, including 3D goggles and an emitter (see Figure 1). A SensAble Technologies PHANTOM Omni® Haptic Device was used as the haptic control interface. The Omni includes a boom-mounted stylus that supports 6 DOF movement and 3 DOF force feedback. The interface recorded subject performance data automatically.



Fig. 1. VR-BD training apparatus including PHANTOM Omni and NVIDIA 3D Vision kit

Experiment sessions were designed to simulate occupational therapy sessions. Two tests, the ROCF and BD subtest from the Wechsler Adult Intelligence Scale – Third Edition (WAIS-III; [14]), represented occupational tasks anticipated to improve as a result of therapy. These two tests were administered once prior to training to evaluate baseline psychomotor performance and again following multiple psychomotor training sessions in order to measure performance improvements. The training task was an updated version of the VR-BD task used in previous studies [5], [12].

The ROCF was administered using a VR adaptation of the task [10]. The task interface was designed to replicate a drawing setup. It included a custom workstation featuring a flat-screen monitor mounted in a tabletop and another Phantom Omni haptic device (see Figure 2, left). To perform the ROCF, subjects used the Omni to virtually draw the complex figure elements directly on the horizontally-aligned monitor (see Figure 2, right, for the ROCF image with numbered units). Rey Osterreith Complex Figure performance is scored by evaluating 18 individual components of the figure that make up a complete design, referred to as units, on a scale from 0 to 2 in terms of accuracy (e.g., size, length) and placement (e.g., proximity to other units). The sum of the scores for the 18 components is calculated for a total score between 0 and 36. The simulation recorded subject test performance data and calculated the scores automatically.

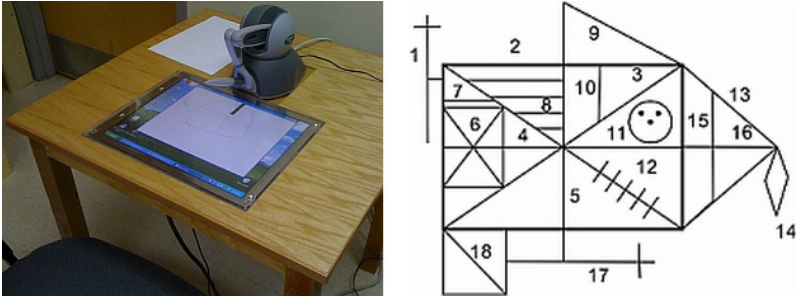


Fig. 2. ROCF workstation and test image

In addition to the ROCF test, subjects completed the WAIS BD subtest during pre- and post-testing to further characterize training effects. The WAIS BD task used for testing and the WASI BD task used for training are identical except for the patterns completed by the subjects. During testing, the WAIS BD was administered using standardized materials.

The features of the VR-BD training task included a virtual tabletop divided into two parts, including a display area (see Figure 3 (a)) and a work area (see Figure 3 (b)). The display area presented the pattern (see Figure 3 (c)) to be replicated by a subject. The work area was used for arranging the blocks. Like the standardized version of the BD task [6], virtual red and white blocks printed with either solid or cross-sectional patterns on each side were distributed randomly in the work area. The design was presented in the display area, and subjects manipulated the blocks to reproduce the picture as quickly as possible. All patterns were constructed with the aid of a target grid (see Figure 3 (d)), which appeared as a 2x2 or 3x3 collection of squares in the work area, depending on the dimensions of the pattern.

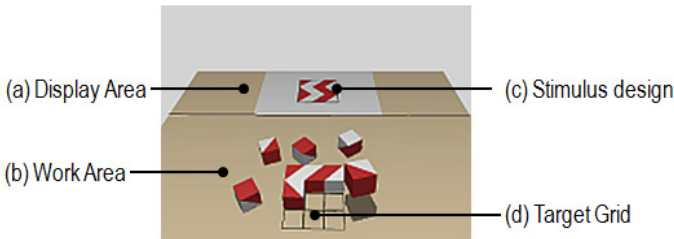


Fig. 3. VR-BD training display layout

The Phantom Omni was used to manipulate a cursor appearing on the display during training. Blocks could be grasped by touching the cursor against them and pressing and releasing the button on the stylus of the haptic device. A block could then be lifted from the table surface and rotated along any axis using the stylus (without holding the button). A block was released upon return to the table surface. Haptic features representing physical properties of the blocks and the table were also included.

The type of aiding represented the independent variable, including the (1) haptic, (2) visual, or (3) combination conditions. The dependent variables included: (1) ROCF test performance and (2) WAIS BD test performance.

2.1 Procedures

There were three main parts of the experiment for data collection: (1) an evaluation of pre-test performance, (2) multiple training sessions, and (3) the post-test to measure improvement. The three parts of the experiment were distributed across four days, with testing scheduled on the first and last days and training taking place on Days 1-3. Each subject completed eight VR-BD trials in total (10 designs per trial, as required by the established WASI protocol). The combined duration of the three training visits was approximately 3 hours, which was established through pilot testing and prior work [5].

The experiment followed a between-subjects design and each subject was assigned to one aiding type (haptic, visual, or combination) for VR-BD training with a total of eight subjects per condition. The combination condition incorporated all the haptic and visual features. Haptic aiding included snap forces that pulled blocks to a target position during correct placement and rejection forces that acted against the block during incorrect placement. Visual aiding provided feedback during incorrect block placements. If a user attempted to place a block in an incorrect orientation in the target grid, a yellow “X” or an arrow would be superimposed on the block. The “X” would appear when the wrong block face was showing (see Figure 4 (a)). The arrow would appear when the correct block face was showing, but it was rotated incorrectly (see Figure 4 (b)). The arrow indicated the direction in which the block needed to be rotated for correct orientation in the target grid square. If a user moved a block in the correct orientation over the grid, those squares in the grid at which the block could be placed without error were highlighted in yellow.

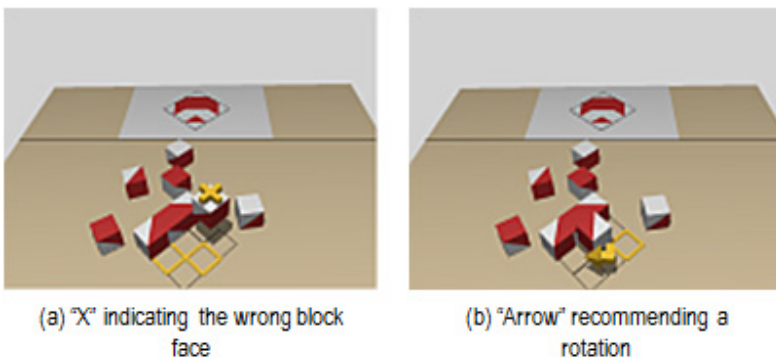


Fig. 3. VR-BD visual assistance during block placement

In addition to the passive visual assistance, subjects could request additional assistance during training. Touching the cursor to the pattern at the top of the screen or the

target grid caused visual cues to be displayed on how to correctly place blocks. Specifically, the cues indicated the orientation and locations of individual block faces. Touching the cursor to a target grid square highlighted the corresponding square in the stimulus pattern and any blocks in the workspace that matched the selected square. Likewise, touching the stimulus pattern would highlight the corresponding square on the target grid. The gridlines disappeared when any surface outside of the stimulus pattern and grid was contacted with the cursor.

2.2 Hypotheses

We hypothesized that all three VR conditions would result in ROCF and WAIS BD test performance improvements (Hypothesis (H1)). Based on the results of previous study [5], it was also expected that training in the combination condition would result in greater improvements in ROCF performance as compared to visual or haptic aiding, alone (H2). This is also consistent with existing notions that suggest receiving feedback via multiple compatible sensory modalities can produce better performance than from a single modality [15].

3 Results

Subject pre-test performance was compared across conditions using Kruskal-Wallis tests. No significant differences attributable to the training condition were revealed. In other words, subjects began data collection at similar performance levels. Pre-and post-test data were analyzed to identify differences in training effects among the three conditions. The results of ROCF and WAIS BD pre- and post-test scores are presented in Table 1.

Table 1. Results of ROCF and WAIS-BD test scores

Condition	N	ROCF			WAIS-BD		
		Pre	Post	%	Pre	Post	%
Comb.	8	25.75	28.25	13.84	44.50	53.00	23.56
Haptic	8	27.00	28.13	7.63	42.50	55.63	31.55
Visual	8	24.63	28.19	18.50	46.38	52.63	16.61

Pre- and post-test scores were compared for each training condition. As a result of some of the response data violating the normality assumption of parametric tests, Wilcoxon rank sum paired tests were conducted to compare the various training conditions. Subject WAIS BD test scores significantly improved as a result of the visual ($p=0.018$), haptic ($p=0.007$) and combination ($p=0.004$) conditions. However, ROCF

test performance did not reveal significant improvements between pre- and post-training (due to a high degree of variability in performance among subjects). Only subjects assigned to the visual condition showed marginally significant improvements in ROCF scores ($p=0.061$). Additional analyses on the percent change in WAIS BD test scores revealed no significant differences in the extent of improvement among training conditions. Statistically speaking, the three conditions resulted in similar increases; however, on average, the haptic condition showed the highest percentage of WAIS BD test improvement.

4 Discussion

The results of the experiment revealed that training in any of the three conditions increased WAIS BD test performance, which was consistent with H1. However, contrary to H1, training did not necessarily lead to increases in ROCF (surrogate occupational task) performance. Beyond this broader result, the degree of improvement in test task performance may vary by condition. There are several possible reasons for this. The snap force feature implemented as part of the haptic aiding condition pulled blocks to their final position as they were moved near the design construction. In effect, while the subject was responsible for gross movement, the honing portion of the task (requiring placement of the block at the target location) was offloaded to the system, and subjects were not required to perform any fine positioning on their own. This means that conditions providing haptic assistance (i.e., the haptic and combination conditions) provided less fine motor skill training than the visual-only aiding condition, which required fine movements during final block positioning. These fine motor skills may have been useful when replicating the ROCF using the haptic device. This may explain the marginally significant increase in ROCF scores under the visual aiding condition (where fine motor movement was not automated) as well as the lack of benefit in terms of ROCF scores as a result of haptic aiding.

Similar results were observed in WAIS BD test performance following training in the visual condition. The visual aiding was designed to assist subjects in parsing the stimulus designs into individual squares corresponding to block faces. This offloaded cognitive aspects of the task to the system; that is, subjects were not required to perform mental segmentation of a block design. There is evidence that subjects receiving visual assistance relied on the automated assistance rather than honing their own cognitive strategies [16]. The additional visual and mental processing of a stimulus pattern required of subjects assigned to the haptic condition likely helped them refine their strategy for stimulus segmentation, as compared to the visually aided subjects. It is likely that this is the reason that haptically aided subjects showed the greatest increases in WAIS BD test performance. While subjects receiving visual assistance were able to rely on visual aids that parsed the design and recommended block orientations and increase scores, subjects receiving haptic assistance had to learn these strategies on their own.

It was expected that training in the combination condition would lead to the greatest increases in test performance due to the presentation of combined haptic and visual

cues (H2). However, results did not support this expectation. In fact, the combination condition led to mediocre training effects in terms of ROCF test score improvement, as compared to the visual-only group, and WAIS BD test score improvement for the haptic-only group. This may be due to the combination of visual and haptic assistance increasing cognitive load or distracting subjects during training. This observation is also consistent with the findings of [7], which proposed that vision may interfere with haptic cues during training.

5 Conclusion

The outcomes of this work are important to VR-based motor training system design. While a form of aiding may be developed to assist psychomotor task performance during training, it may also hinder development of motor and cognitive skill requirements that are allocated to automated assistance. This raises a distinction between designing for training task performance and designing for motor skill learning. During VR-BD task training, subjects could rely on visual aiding instead of developing a strategy for parsing the blocks in the model [16]. However, by offloading some cognitive aspects of the task to the automation, these subjects received less training that could improve WAIS BD test scores where visual aiding was not available.

One limitation of the present study was the use of unimpaired subjects. Although parallels were drawn between physical and cognitive characteristics of non-dominant performance and motor planning and control implications of mTBI, there is a need to test an actual pathological population using the VR technology. For the next phase of this research, in addition to recruiting unimpaired subjects, we will recruit subjects from a pathological population to observe the effects of VR-based haptic training on patients with a history of mTBI, including fine motor skill implications. We also plan to extend the test data by incorporating functional magnetic resonance imaging (fMRI) during pre- and post-test procedures to measure changes in brain activity as a result of VR-BD training.

Acknowledgements. This research was supported by a grant from the National Science Foundation (NSF) (No. IIS-0905505) to North Carolina State University. The technical monitor was Ephraim Glinert. The views and opinions expressed on all pages and in all documents are those of the authors and do not necessarily reflect the views of the NSF.

References

1. Merians, A.S.: Virtual Reality-Augmented Rehabilitation for Patients Following Stroke. *Physical Therapy* 82(9), 898 (2002)
2. Wiederhold, B.K., Wiederhold, M.D.: The future of cybertherapy: improved options with advanced technologies. *Studies in Health Technology and Informatics* 99, 263–270 (2004)

3. Jang, S.H., You, S.H., Hallett, M., Cho, Y.W., Park, C.M., Cho, S.H., et al.: Cortical reorganization and associated functional motor recovery after virtual reality in patients with chronic stroke: an experimenter-blind preliminary study. *Archives of Physical Medicine and Rehabilitation* 86(11), 2218–2223 (2005)
4. Holden, M.K.: Virtual environments for motor rehabilitation: review. *Cyberpsychology & Behavior* 8(3), 187–211 (2005)
5. Clamann, M., Gil, G.-H., Kaber, D.B., Zhu, B., Swangnetr, M., Jeon, W., Zhang, Y., Qin, X., Ma, W., Tupler, L.A., Lee, Y.-S.: Assessment of a virtual reality-based haptic simulation for motor skill rehabilitation. In: *Proceedings of the 2012 Applied Human Factors & Ergonomics Conference (CD-ROM)*. Taylor & Francis, CRC Press, Boca Raton, FL (2012)
6. PsychCorp: Wechsler Abbreviated Scale of Intelligence (WASI) manual. Pearson Education, Inc. (1999)
7. Feygin, D., Keehner, M., Tendick, R.: Haptic guidance: experimental evaluation of a haptic training method for a perceptual motor skill. In: *Proceedings of 10th Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems (HAPTICS 2002)*, Orlando, FL, pp. 40–47 (2002)
8. Liu, J., Cramer, S.C., Reinkensmeyer, D.J.: Learning to perform a new movement with robotic assistance: comparison of haptic guidance and visual demonstration. *Journal of NeuroEngineering and Rehabilitation* 3(20) (2006)
9. Osterreith, P.A.: Le test de copie d'une figure complexe (The complex figure copy test). *Archives de Psychologie* 30, 206–356 (1944)
10. Li, Y., Kaber, D.B., Lee, Y.-S., Tupler, L.: Haptic-based virtual environment design and modeling of motor skill assessment for brain injury patients rehabilitation. *Computer-Aided Design and Applications* 8(2), 149–162 (2010)
11. Basdogan, C., Kiraz, A., Bukusoglu, I., Varol, A., Doğanay, S.: Haptic guidance for improved task performance in steering microparticles with optical tweezers. *Optics Express* 15(18), 11616–11621 (2007)
12. Jeon, W., Clamann, M., Zhu, B., Gil, G.-H., Kaber, D.B.: Usability evaluation of a virtual reality system for motor skill training. In: *Proceedings of the 2012 Applied Human Factors & Ergonomics Conference*, Boca Raton, FL (2012)
13. Oldfield, R.C.: The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia* 9, 97–114 (1971)
14. Wechsler, D.: WAIS-III administration and scoring manual. The Psychological Corporation, SanAntonio (1997)
15. Ernst, M., Banks, M.: Does vision always dominate haptics? In: *Touch in Virtual Environments: Haptics and the Design of Interactive Systems*. Prentice Hall, Upper Saddle River (2002)
16. Clamann, M., Kaber, D.B.: The Effects of Haptic and Visual Aiding on Psychomotor Task Strategy Development During Virtual Reality-Based Training. In: *Proceedings of the Human Factors and Ergonomics Society annual Meeting (CD-ROM)* (2012)

Influence of Haptic Feedback on a Pointing Task in a Haptically Enhanced 3D Virtual Environment

Brendan Corbett¹, Takehiko Yamaguchi², Shijing Liu¹,
Lixiao Huang¹, Sangwoo Bahn¹, and Chang S. Nam¹

¹ North Carolina State University, Raleigh, NC, USA

{bcorbet, sliu14, lhuang11, csnam}@ncsu.edu, panlot@gmail.com

² Université d'Angers, Angers, France

t.yama007@gmail.com

Abstract. To gain a better view of the value of haptic feedback, human performance and preference in a pointing style task in a three-dimensional virtual environment was explored. Vibration and haptic attractive force were selected as two simple cases of feedback, each with two levels. These types of feedback were compared to a no-feedback condition to better understand how human performance changes under these conditions. The study included 8 undergraduate students. A Novint Falcon haptic controller was used in a simulated three-dimensional virtual environment. Analysis was conducted on how each type of feedback effects the movement time (MT) of users. The results showed that vibration was perceived negatively and had a slight negative impact on performance. The haptic attractive force significantly improved performance and was strongly preferred by subjects.

Keywords: Haptic, assistive technology, virtual environments, human performance, force feedback, vibration, assistive feedback.

1 Introduction

Haptic feedback is a rapidly growing research emphasis in human computer interaction. The ability to utilize the somatic sense, or sense of touch, can greatly enhance the realism and improve immersion, provide additional awareness through redundancy, or provide assistive support (Robles-De-La-Torre, 2006). Identifying subject preference for types of feedback and performance effects of haptic assistive feedback indicated potential value in functionally similar tasks. For example, a visually impaired person may be able to utilize assistive feedback to better navigate a virtual environment, improving their computer interaction experience. A variety of studies have utilized haptic feedback in medical applications, such as dentist training (Suebnuakarn, et al., 2009) and using a Leksell Gamma Knife to neutralize tumors in the brain (Dinka, Nyce, Timpka, & Holmberg, 2006). In these studies, participants' performance and attitudes toward these haptic feedbacks varied. Further research is needed to determine whether haptic feedback improves performance in an objective, general case. To examine the effect of haptic assistive feedback in a simple, performance based Fitts' Law style task was implemented. Fitts (1954) initially

proposed a one-dimensional tapping task, with subjects alternating the tapping between two targets. This task resulted in the following equation being derived:

$$MT = a + b * \log_2(A/W) \quad (1)$$

Where MT is the movement time from a start point to a target, A is the amplitude of distance from start to the target, and W is the width of the target. The values of a and b are constants calculated from experiment data. The logarithm portion is commonly referred to as the Index of Difficulty (ID). The simple linear relationship between ID and MT has been applied in a wide variety of studies. Furthermore, Fitts identified the Index of Performance (IP) as $1/b$ in bits/s. This relationship between amplitude and width has been successfully extended into two-dimensional computer environments as well by Accot and Zhai (2003). Further studies have shown results consistent with the original Fitts equation in three dimensions and with different controllers including haptic control devices (Murata, 2001, Mateo et. al., 2005, Campbell et. al., 2008, Margolis et. al, 2011).

To study the effect of haptic assistive feedback on human performance, we had subjects complete a basic Fitts' style task in a three dimensional virtual environment. Subjects would move their cursor from a start point to a target object in the virtual environment, then click to activate it. This would be conducted with no haptic feedback, vibration feedback, and a haptic attractive force feedback.

2 Methods

2.1 Participants

This study included eight undergraduate students as participants, six male and two female, ranging in age from 18 to 20. None of subjects had prior experience with haptic virtual environments.

2.2 System Design

Subjects utilized a Novint Falcon haptic controller to interact with a three dimensional virtual environment. The system was developed based on the Novint Falcon SDK

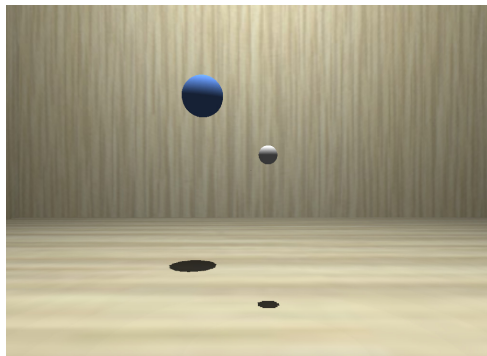


Fig. 1. Subject view of experiment virtual environment

specifically for this study. The subject view of the virtual environment can be seen in Figure 1.

The environment provided a view of looking into a workspace similar to a box. The space was scaled to fit just within the maximum motion range of the Novint Falcon device. Targets would appear in random locations in the workspace. The target size would be sampled from a matrix with three possible sizes. Similarly, the distance to target was sampled from a matrix with three possible distances, and a random location was selected based on the distance vector.

2.3 Feedback

The control condition employed no haptic feedback. Users were simply tasked with moving their cursor as quickly and accurately to the target object, then clicking.

Vibration feedback was designed to improve the transition from ballistic to homing motion. Users feel a constant, light vibration in the controller until they are within close proximity to the target. Two levels of vibration feedback were employed.

Attractive haptic force feedback provided a moderate attractive force from the users' cursor to the target, regardless of cursor position. The construct used was similar to a spring, providing a positive addition to user input force in the direction of the target. The feedback itself would not automatically move the cursor to the target, rather it would amplify user-initiated motion. Two levels of attractive haptic force feedback were employed.

2.4 Procedure

Subjects initially completed a brief training session to familiarize them with the virtual environment, the haptic controller, and the types of feedback they would experience. Following the training session, subjects completed the main trials.

Each trial employed the same basic task, based on extending Fitts' Law to three dimensions. At the beginning of each trial, a large white sphere, or start object, was present in the interface. Users would move their cursor to the start object and click. Once they had clicked the object, the start object disappeared and a blue sphere, or target object, would appear. The objective was to move the cursor to the target object as quickly and accurately as possible, then click the center button on the top of the Novint Falcon. The target then disappeared and a new target appeared. Movement time, the primary response measure, was recorded from the time the start object or target was clicked to the time the next target was clicked.

As previously stated, three target widths and three amplitudes of distance to target were used, creating nine unique width-distance pairs. Each pair was replicated three times in a trial, resulting 27 total target objects in each trial. The feedback in each trial was provided in a random order. Aside from feedback, the main trials were identical.

Following the main trials, subjects completed a survey to rank their preference for feedback and assessed their perceived performance under each feedback type.

3 Results and Discussion

The results for Fitts' parameters can be seen in Table 1, grouped by feedback type.

Table 1. Fitts parameters a , b , and IP , by feedback type

	Haptic	Vibration	None
a	0.57	-0.02	0.35
b	0.33	0.56	0.46
IP	3.00	1.77	2.20

The Index of Performance had clear implications about the effect of the feedback on performance. The significantly higher IP of 3.00 bits/s for the haptic attractive force feedback significantly outweighs the IP of 1.77 bits/s for vibration feedback. While the addition of assistive feedback of any time would presumably improve performance, the IP for vibration feedback was actually lower than that of the no feedback trials.

A plot of the average movement time for each given ID provides more insight into the relationship. Figure 2 presents a plot of average MT by ID for each type of feedback.

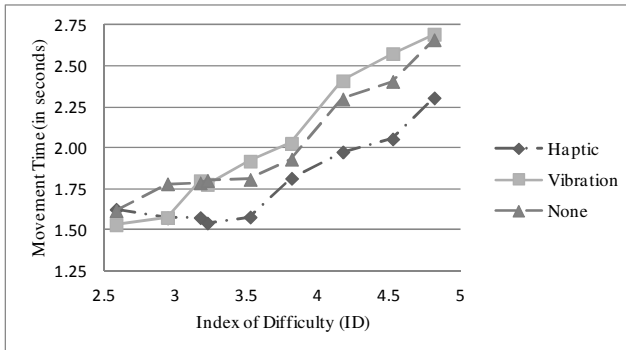


Fig. 2. Plot of average movement time by index of difficulty for each feedback type

Of interest are the data points for lower IDs. The overlap of the three data series implies that the task difficulty at these levels was low enough to not generate different user behavior under the conditions, meaning users could likely complete the task purely with ballistic motion. Furthermore, the coefficients of determination, or r -square values, for the linear regression are 0.976 for vibration, 0.923 for no feedback, but only 0.837 for haptic feedback. To analyze only the data for which the ID is sufficient to require both ballistic and homing motion, the analysis is repeated after removing the lowest 3 IDs. Table 2 contains the adjusted Fitts' parameters a , b , and IP .

Table 2. Adjusted Fitts parameters a, b, and IP, by feedback type for highest 6 IDs

	Haptic	Vibration	None
a	-0.03	-0.23	-0.14
b	0.48	0.61	0.57
IP	2.10	1.63	1.75

The adjusted values show a similar relationship between feedback types. The IP for haptic attractive force feedback still remains the highest at 2.10 bits/s, no feedback at 1.75 bits/s, and vibration indicated performance below the no feedback condition at 1.63 bits/s. The adjust plot can be seen in Figure 3.

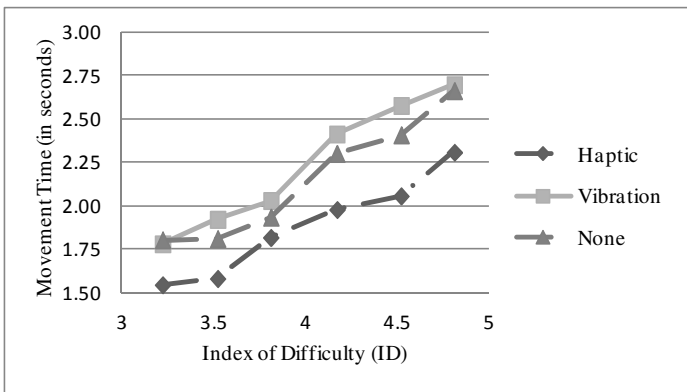


Fig. 3. Adjusted movement time by index of difficulty for each feedback type

The coefficients of determination indicated a much better fit. For haptic attractive force feedback, the value is 0.967, for vibration 0.978, and for no feedback 0.949. The linear regressions were a better fit for this set of IDs. With a sufficiently high ID to require both ballistic and homing motion, human performance in a three dimensional virtual environment conforms to the parameters of Fitts' Law. Furthermore, haptic attractive force indicates an improved performance. Because of the implementation of the attractive force feedback, the constant amplification of motion towards the target should minimize the need for error correction, increasing efficiency in error correction and improving the accuracy of ballistic motion. By improving the alignment and accuracy of ballistic motion, less time should be required for fine homing motion to touch the target, reducing overall movement time.

To complement the performance results, subjects were asked to self-rate their performance by feedback type. Figure 4 shows the results of user self-rating of performance.

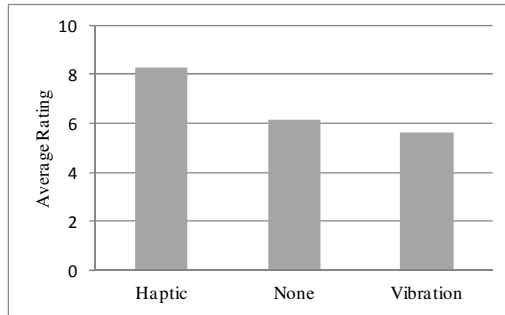


Fig. 4. Self rating of performance by feedback type

These results closely match performance results. Subjects believed their performance to be significantly better with the haptic attractive force. This is the expected outcome for assistive feedback. The goal of providing additional assistive feedback was to both improve performance and improve a user's experience. The vibration results are consistent with previous research on human response to vibration. The sense of urgency is elicited from the feedback. Regardless of whether the vibration conceptually would help, the human reaction is innately negative, possibly influencing the end performance.

Subjects were also asked to rank their preferred feedback, as seen in Figure 5.

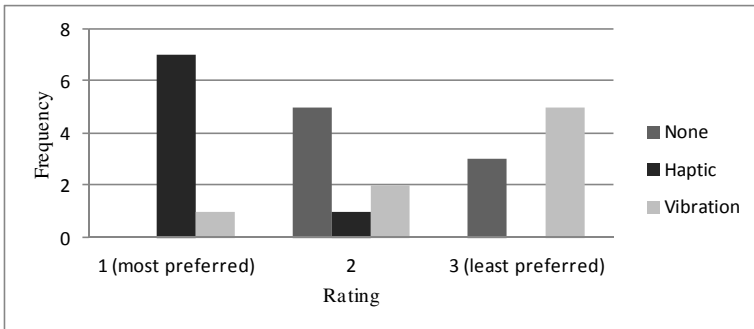


Fig. 5. Frequency count of user preference ranking

The subjective ranking indicates a strong preference for haptic feedback. Similar to performance and self-assessment of performance, subjects showed a lack of interest in vibration feedback, again related to the natural sense of alarm induced by vibration.

4 Conclusion

Subjects completed a pointing task similar to the original Fitts task, in a three dimensional virtual environment with and without haptic feedback. Results indicate

that Fitts' Law holds true for this scenario, with a linear relationship between ID and MT. Haptic attractive force feedback elicited the best performance results, along with high user ranking in preference. Vibration feedback resulted in performance similar to no feedback, however user preference was generally low for this feedback. Subject self-assessment of performance was consistent with actual performance results. Ultimately haptic attractive force feedback has a positive effect on performance and users also indicate a preference for the feedback.

Further studies may look at a broader scope of haptic feedback to generate a more clear depiction of human performance. There may be other scenarios or ways to implement vibration feedback that positively influence performance and improve user perception. Haptic feedback may also have more applicability in different virtual environments where the additional multimodal feedback may have a greater effect.

References

1. Accot, J., Zhai, S.: Refining Fitts' law models for bivariate pointing. In: Proceedings of ACM Conference on Human Factors in Computing Systems - CHI 2003, pp. 193–200 (2003)
2. Campbell, B.A., O'Brien, K.R., Byrne, M.D., Bachman, B.J.: Fitts' law predictions with an alternative pointing device (wiimote®). In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 52(19), pp. 1321–1325. SAGE Publications (2008)
3. Dinka, D., Nyce, J.M., Timpka, T., Holmberg, K.: Adding value with 3D visualization and haptic forces to radiosurgery – A small theory based, quasi-experimental study. *Journal of Medical Systems* 30, 293–301 (2006)
4. Fitts, P.: The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology* 47(6), 381–391 (1954)
5. Margolis, T., DeFanti, T.A., Dawe, G., Prudhomme, A., Schulze, J.P., Cutchin, S.: Low cost heads-up virtual reality (HUVR) with optical tracking and haptic feedback. *IS & T/SPIE Electronic Imaging* 0001, 786417-786417-11 (2011)
6. Mateo, J.C., Manning, J.T., Cowgill, J.L., Moore, T.J., Gilkey, R.H., Simpson, B., Weisenberger, J.M.: Evaluation of a collaborative movement task in a distributed three-dimensional virtual environment. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 49, pp. 1578–1582 (2005)
7. Murata, A., Iwase, H.: Extending Fitts' law to a three-dimensional pointing task. *Human Movement Science* 20(6), 791–805 (2001)
8. Robles-De-La-Torre, G.: The importance of the sense of touch in virtual and real environments. In: *Haptic User Interface for Multimedia Systems*, pp. 24–30 (July–September 2006)
9. Suebnukarn, S., Phatthanasathiankul, N., Sombatweroje, S., Rhiennora, P., Haddawy, P.: Process and outcome measures of expert/novice performance on a haptic virtual reality system. *Journal of Dentistry* 37, 658–665 (2009)

Design of a Wearable Haptic Vest as a Supportive Tool for Navigation

Anak Agung Gede Dharma¹, Takuma Oami¹,
Yuhki Obata¹, Yan Li¹, and Kiyoshi Tomimatsu²

¹ Kyushu University, Graduate School of Design, Fukuoka, Japan
dharma@kyudai.jp

² Kyushu University, Faculty of Design, Fukuoka, Japan
tomimatsu@design.kyushu-u.ac.jp

Abstract. We propose an alternative way to display haptic feedback in ubiquitous computing. We develop a haptic vest that can display detailed haptic feedbacks by utilizing 5x12 arrays of vibrotactile actuators. We conducted a preliminary user testing on 34 stimuli (with four different directions) to measure the effectiveness of various vibrotactile patterns. We have discovered that each stimulus within a given direction has different properties in terms of their apprehensibility and comfort.

Keywords: Wearable computing, haptic rendering, haptic perception.

1 Introduction

In recent years, haptic feedback has been intensively researched as the next generation of multimodal human computer interaction. The user experience that revolves around haptic feedback is not only limited to the material properties, such as softness or hardness, the recent advance in information technology also makes it applicable to be used in various forms. Furthermore, haptic feedback plays indispensable roles in several scenarios due to its unique characteristics, such as the capability of delivering private information in a non-intrusive way. Therefore, it can be applied on numerous purposes such as wayfinding [1] or therapy [2].

However, to fully utilize its potential, it is important to explore the spatial and temporal characteristics in designing haptic devices. Therefore, we propose the design of a wearable haptic vest as a supportive tool in navigation in this study.

This paper mainly explores the user experience of interacting with the haptic vest as a navigation system. The basic design concept of our proposed haptic vest is the idea that users get the information regarding direction by sensing vibrations from 5x12 arrays of actuators. The main consideration is to explore the most effective way to convey the directional information to users. Furthermore, the comprehensive understanding regarding this phenomenon could be used to propose a better solution for users on interacting with haptic vest or other haptic wearable devices in general.

2 Related Works

The psychophysical sensation of feeling tactile illusions according to two-dimensional tactile actuation has been reported by Israr et al. [3, 4]. Furthermore, preceding works by Vaucelle et al. [2], Ertan et al. [5], and Tan et al. [6] suggest that an $n \times n$ array of actuators can be used to haptically convey information to the users. However, these researches only utilized a relatively small number of arrays, i.e. 3×3 [6] or 4×4 [5]. To gain better insight at the haptic information, we consider that more navigation patterns are needed, which can be achieved by increasing the number of actuators array.

We propose a haptic vest that consists of 5×12 arrays of actuators. By connecting these actuators to a microcontroller, it is possible to independently control each actuator and create sophisticated design patterns with various spatial and temporal characteristics. Another aspect that we thoroughly consider during the design process is the vest material. We chose a wet suit as the main material as it provides sufficient strength, durability, and elasticity. In addition, we have conducted a preliminary user testing to ensure the wet suit is suitable for our experiment. The design concept of our proposed haptic vest is described in Figure 1.



Fig. 1. The design concept of our proposed wearable haptic vest, the left figure and right figure show its outer and inner view, respectively

3 Research Objective

The main objective of this study is to explore kinds of user experience that are associated when interacting with the haptic vest. We have created various haptic patterns to test how users feel the haptic stimulus and perceive information from it. The goal is to discover better ways to utilize haptic patterns in navigation systems.

The actuators on the haptic vest are controlled by Arduino¹ that receive navigation information from Central Processing Unit (CPU) and convert it into vibration pattern

¹ An open source prototyping platform allowing to create interactive electronic objects
<http://www.arduino.cc>

within 5x12 arrays of actuators. Users sense the haptic information when they come in contact with the vibration. Figure 1 shows the outer and inner view of the haptic vest. However, when worn, the actuators are within the inner side of the haptic vest.

4 Research Method

4.1 Participants

Sixteen subjects (7 males and 9 females) participated in the experiment. All subjects are undergraduate students from School of Design, Kyushu University.

4.2 Experiment Setting

Our experiment subject wears the haptic vest as shown in Figure 2. The user testing experiment is done by the following sequence:

1. The subject is asked to wear the haptic vest and sit throughout the experiment.
2. A stimulus is chosen randomly and exposed to the subject. The stimulus is being played in a continuous loop until the subject gives a response.
3. The subject was asked to choose between four directions (back, front, left, and right) based on his/her perceived direction of the haptic stimulus.
4. After giving a guess on the direction, the subject were told whether the stimulus point the same direction as his/her guess or a different one.
5. The subject were asked to rate the stimulus' comfort and apprehensibility on a five step Likert scale.
6. The process is repeated from step (2) until the last stimulus has been exposed.



Fig. 2. Experiment setting for the usability test

The experiment is done within one hour for each subject. Before the real experiment begins, users were subjected to “a training session” by guessing and giving

subjective evaluations (on comfort and apprehensibility) on one random dummy stimulus. In this experiment, we used FM34F² as the vibrotactile actuators.

4.3 Haptic Patterns

Our stimuli for the user testing experiment consist of 34 unique haptic patterns, each indicate one of four main directions (back, front, left, or right). These patterns include eight front patterns, eight back patterns, nine left patterns, and nine right patterns. Each stimulus varies on the vibration strength, haptic pattern, and the total exposure time (Figure 4, 5, 6, and 7). The levels of vibration strength in those figures are described by color contrasts, as illustrated in Figure 3. We prepared fourteen different levels of vibration strength. The strengths are proportionally increased as a linear function. For instance, level 2 is twice as strong as level 1 and level 3 is three times as strong as level 1.

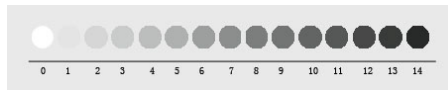


Fig. 3. Fourteen different levels of vibration strength

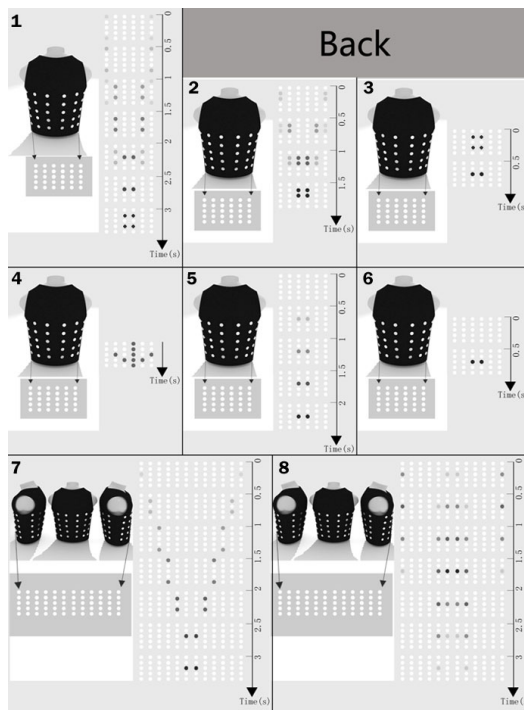


Fig. 4. Haptic patterns for the back direction

² Small disk-shaped vibration motor that is manufactured by T.P.C.

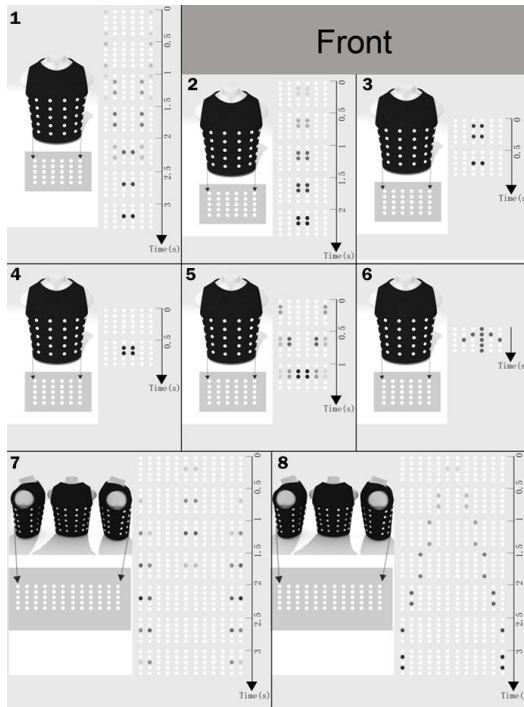


Fig. 5. Haptic patterns for the front direction

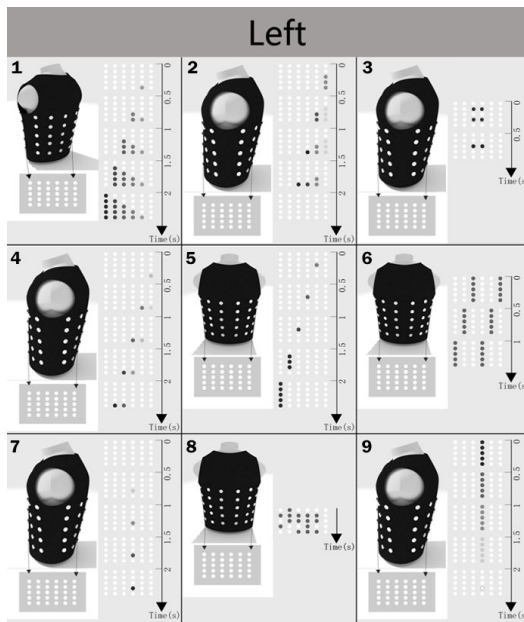


Fig. 6. Haptic patterns for the left direction

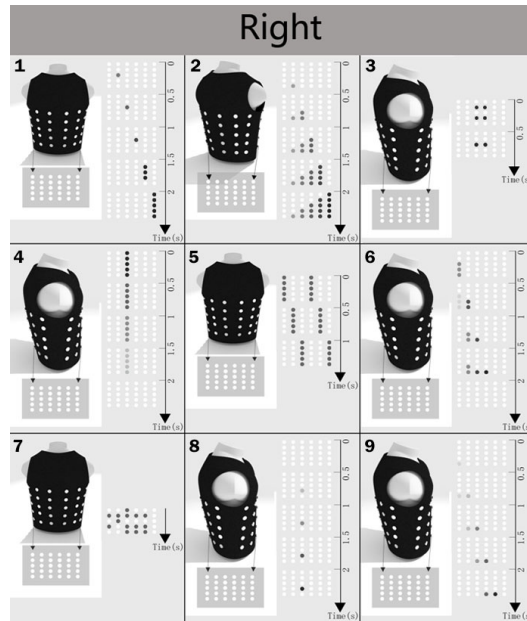


Fig. 7. Haptic patterns for the right direction

To obtain a comprehensive understanding of perceived haptic sensation to the upper torso, various types of patterns are tested in this experiment. The patterns are varied from simple static binary patterns [e.g., Back-6 (B6) and Front-4 (F4)], static gradation patterns [e.g., Left-9 (L9) and Right-4 (R4)], to dynamic gradation patterns (e.g., L1 and R2). In addition, for the lateral patterns, we prepared two patterns that resemble Japanese character for left (L8) and right (R7).

5 Experiment Results

The result of the user testing experiment is summarized in Table 1 and visualized in Figure 8. It can be concluded that haptic patterns within the top-right corner of Figure 8 are preferred by users, due to their high ratings of comfort and apprehensibility.

For the left direction, relatively simple patterns (e.g., L2, L7, and L9) are preferred to complicated patterns (e.g., L5, L6, and L8). The similar finding of their mirrored version can be observed for the right direction. In this direction, simple patterns are also preferred (e.g., R4, R6, R8) to complicated right patterns (e.g., R1, R5, R7). However, the patterns that resemble Japanese characters for left (L8) and right (R7) are not preferred by users and have low apprehensibility ratings.

For the back direction, although almost all haptic patterns have relatively high ratings of comfort and apprehensibility (with the exception of B8), simple patterns like B2, B3, and B6 are preferred by users. However, for the front direction, all haptic patterns have low rating on comfort. Simple patterns like F2, F3, and F4 can be described as preferable patterns due to their high ratings on apprehensibility.

Table 1. Corresponding directions, percentage of correct scores, apprehensibility, and comfort of each haptic pattern

Haptic Pattern*	Direction	Correct Scores (%)**	Apprehensibility (average)	Apprehensibility (st. dev.)	Comfort (average)	Comfort (st. dev.)
B1	Back	93.75	3.312	0.946	3.562	1.094
B2	Back	100	4.125	0.885	3.562	0.892
B3	Back	100	4.312	0.873	3.562	0.964
B4	Back	87.5	3.312	1.352	3.625	0.957
B5	Back	100	3.75	0.856	3.625	0.885
B6	Back	100	4.5	0.73	3.687	0.873
B7	Back	100	2.5	1.155	3.125	1.025
B8	Back	93.75	2.875	1.455	3.062	0.854
F1	Front	100	3.625	0.957	2.75	0.856
F2	Front	100	4.562	0.727	2.562	1.094
F3	Front	100	4.187	1.047	2.625	1.025
F4	Front	100	4.437	1.153	2.5	0.966
F5	Front	100	3.75	0.856	2.625	1.025
F6	Front	100	4.062	1.237	2.312	1.014
F7	Front	100	3.625	1.204	2.625	0.719
F8	Front	100	3.125	0.957	2.5	1.095
L1	Left	75	3.187	1.167	3.312	0.873
L2	Left	87.5	3.187	1.109	3.562	1.031
L3	Left	100	4.625	0.619	2.312	1.195
L4	Left	100	3.375	0.719	2.875	0.619
L5	Left	81.25	2.5	1.033	3.062	0.854
L6	Left	6.25	1.5	0.632	3.25	1.065
L7	Left	100	4	0.894	3	1.155
L8	Left	6.25	1.5	0.632	3.187	1.109
L9	Left	100	4.5	0.816	2.812	0.981
R1	Right	100	2.625	1.088	2.812	0.75
R2	Right	100	3.062	1.289	3.187	0.911
R3	Right	100	4.625	0.619	2.687	1.138
R4	Right	100	4.187	0.981	3.125	0.957
R5	Right	12.5	1.687	0.704	3.125	1.258

Table 2. (Continued)

Haptic Pattern*	Direction	Correct Scores (%)**	Apprehensibility (average)	Apprehensibility (st. dev.)	Comfort (average)	Comfort (st. dev.)
R6	Right	100	3	1.155	3.375	0.957
R7	Right	0	1.5	1.033	3	1.095
R8	Right	100	4	0.966	3	0.894
R9	Right	100	3.062	1.237	2.875	0.957

* Haptic patterns that are tested in our experiment. B, F, L, and R means back, front, left, and right pattern, respectively.

** The number of users who correctly guessed the direction of the haptic pattern (in percentage)

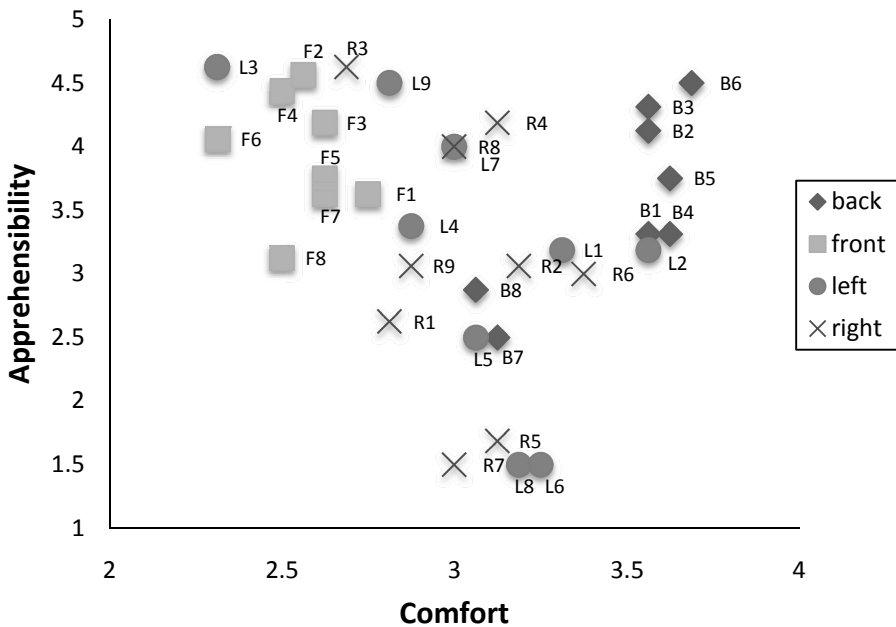


Fig. 8. Comfort vs. Apprehensibility plot for 34 haptic patterns

6 Discussion

From Figure 8, principal characteristics regarding users’ response to haptic stimulation can be analyzed thoroughly. We consider that both of apprehensibility and comfort have to be taken into account when designing haptic patterns for navigation or guidance.

As described in section 5, front vibrations generally elicit uncomfortable sensation, although they are relatively easy to apprehend. Furthermore, among all of the four directions, our subjects report that back vibrations are the most comfortable.

The ratings for haptic patterns for lateral directions (left and right) vary from “very hard to apprehend” to “very easy to apprehend.” On the other hand, lateral vibrations generally have medium level of comfort. For lateral directions, simple haptic patterns are easy to apprehend and perceived to be comfortable, such as R4, R8, L7, and L9. However, there are exceptions, i.e., some simple haptic patterns are evaluated to be uncomfortable although they have high apprehensibility ratings (e.g., L3 and R3). On the other hand, users rate complicated patterns, such as those that resembles Japanese characters (i.e., L8 and R7) to be difficult to apprehend. This finding suggests that users do not have enough capabilities do distinguish spatial properties when given haptic stimulation on their back.

Haptic patterns for the back direction are generally rated to be comfortable and can be apprehend effortlessly. However, B8 and B7 scores relatively low on both comfort and apprehensibility.

Furthermore, haptic patterns for the front direction have low scores on comfort. Therefore, the proposition to determine preferable patterns for this direction should be made only based on their apprehensibility.

From these haptic patterns analysis, we can suggest that spatial and temporal properties of the vibration’s strength play an important role in haptic navigation system. There are notable findings that should be considered when these patterns are adapted for navigation. The lateral directions may use similar types of vibration (i.e., the mirrored patterns), while the forward and back directions should utilize distinct vibration types. Furthermore, when designing the vibration type of going forward, some users report that they may miscomprehend it as the signal that tells him/her to stop instead of going forward, due to the discomfort.

Furthermore, we have observed that different parts of the body experience the haptic stimulation in different ways. In general, the area around the waist is the most sensitive part of the body. Furthermore, when given the same strength of vibration, the front one is perceived to be stronger and more uncomfortable compared to the back one. Therefore, the best way to make users effortlessly apprehend the haptic pattern is to keep it simple and localized within a certain part of the body.

7 Conclusion and Future Works

In this paper, we have proposed 5x12 arrays of vibrotactile actuators that are installed inside the haptic vest as a possible method of displaying sophisticated haptic stimuli. Furthermore, according to the user testing, basic characteristics of perceived haptic stimulation to the users, such as comfort and apprehensibility, have been explored.

We have confirmed that users evaluate back vibrations to be the most comfortable one, while front vibrations are generally evaluated as uncomfortable. Users generally preferred simple haptic patterns to complicated ones, and rate those patterns highly both in terms of comfort and apprehensibility. However, there are some exceptions to

this finding, such as simple haptic patterns that have high vibration strength. Those patterns typically have low score of comfort, although they are still rated to be highly apprehensible.

Future works will include the installation of wireless technology to the haptic vest, application development, and the improvement of its user interface.

Acknowledgements. This work was supported by JSPS KAKENHI Grant Number 245436. Furthermore, we would like to express appreciation to the following individuals for the contributions they made in this study: Marina Ishikawa, Miyuki Kumagai, Moeki Ohwaki, and Natsuki Fukami.

References

1. Heuten, W., Henze, N., Boll, S., Pielot, M.: Tactile wayfinder: a non-visual support system for wayfinding. In: Proceedings of the 5th Nordic Conference on Human-Computer Interaction: Building Bridges, Lund, Sweden, pp. 172–181 (October 2008)
2. Vaucelle, C., Bonann, L., Ishii, H.: Design of haptic interfaces for therapy. In: Proc. 27th International Conference on Human Factors in Computing Systems, Boston, USA (April 2009)
3. Israr, A., Poupyrev, I.: Control space of apparent haptic motion. In: Proc. IEEE World Haptics Conference 2011, Istanbul, Turkey, pp. 457–462 (June 2011)
4. Israr, A., Poupyrev, I.: Tactile brush: Drawing on skin with a tactile grid display. In: Proc. 2011 Annual Conference on Human Factors in Computing Systems, Vancouver, Canada, pp. 2019–2028 (May 2011)
5. Ertan, S., Lee, C., Willets, A., Tan, H., Pentland, A.: A wearable haptic navigation guidance system. In: Digest of Second International Symposium on Wearable Computers, Pittsburgh, USA, pp. 164–165 (October 1998)
6. Tan, H.Z., Gray, R., Young, J.J., Traylor, R.: A haptic display for attentional and directional cueing. *Haptics-e: The Electronic Journal of Haptic Research* 1(3) (2003)

Mapping Texture Phase Diagram of Artificial Haptic Stimuli Generated by Vibrotactile Actuators

Anak Agung Gede Dharma¹ and Kiyoshi Tomimatsu²

¹ Kyushu University, Graduate School of Design, Fukuoka, Japan
dharma@kyudai.jp

² Kyushu University, Faculty of Design, Fukuoka, Japan
tomimatu@design.kyushu-u.ac.jp

Abstract. We propose a classification method of tactile sensations elicited by artificial haptic stimuli by using Japanese onomatopoeias/adjectives. This method classifies adjectives based on user subjective perception and plot basic components of artificial haptic stimuli. The comparison of perceived tactile sensations from artificial haptic stimuli and genuine physical materials is also discussed in this paper.

Keywords: Touch perception, artificial haptic stimuli, Japanese onomatopoeia, Principal Component Analysis.

1 Introduction

In this age of media and telecommunication, haptic feedback plays an indispensable role. Haptic feedback is often used simultaneously with visual or audio feedback. However, when separately used, it still has a capability to convey non-intrusive messages. Notable examples of its importance can be found in various devices, such as mobile phone, Personal Digital Assistant (PDA), game controllers, and medical instruments. Furthermore, haptic feedback also plays a significant role for supporting daily lives of visually impaired persons.

On the other hand, haptic feedback design still has remaining problems. Even in this age of media and telecommunication, it is still limited to a combination of simple force patterns. The study about haptic perception is still limited, especially the study regarding sensations elicited by artificial haptic stimuli.

This study proposes a classification method for explaining tactile sensation generated by artificial haptic stimuli. In this study, we evaluate 100 randomly generated artificial vibrotactile stimuli with subjective evaluation method (Semantic Differential Test) and develop a texture phase diagram using the result from Principle Component Analysis. In the sixth section, detailed analysis of the texture phase diagram and its comparison with preceding researches will also be discussed.

2 Related Works

The attempts of interpreting the correlations between haptic perceptions and physical properties of materials have been discussed in previous studies. Chen et al. [1] and Shirado et al. [2] suggested correlation model to explain the relationships between touch perception and surface physical properties. Shirado et al. [3] proposed a modeling of tactile texture recognition mechanism using human subjective evaluation and computer simulation based recognition model. Furthermore, Hollins et al. [4] and Tiest et al. [5] utilizes Multi Dimensional Scaling to analyze and explain haptic perceptions.

A method to classify tactile textures by using Japanese onomatopoeia has been proposed by Hayakawa et al. [6]. They used 42 Japanese onomatopoeias and Semantic Differential Test to develop a distribution diagram that can explain the correlation between tactile perception and physical characteristics of materials.

3 Research Objective

The main objective of this study is to explore the correlation model of tactile perceptions that are induced by artificial vibrotactile stimuli. The correlation model can be used to determine cumulative effect that artificial vibrotactile stimuli contribute towards overall tactile perception and relative importance of each extracted components. The cumulative effect can determine whether artificial vibrotactile can be used as an adequate replacement of genuine physical materials or not.

4 Experiment Method

4.1 Participants

Fourteen subjects (7 males and 7 females; mean and standard deviation of age were 21.4 and 1.4, respectively) participated in the experiment. All subjects are native speakers of Japanese.

4.2 Stimuli

Selective Stimulation Method

The theory behind selective stimulation method is that tactile receptors in the human skin cannot sense physical factors directly. They can only detect inner skin deformation caused by contact with objects [7]. Therefore, we may be able to use artificial tactile stimulation to activate tactile receptors' nerves as if they were being activated by physical properties of tangible material.

The selective stimulation method is based on the direct manipulation of three tactile receptors: Fast Adapting Afferents Type I (FA1), Fast Adapting Afferents Type II (FA2), and Slow Adapting Afferents Type I (SA1). Each receptor has spatial and temporal response characteristics for physical stimulation, as described in Figure 1. Each receptor also causes subjective sensation that corresponds to inner skin deformation.

In this study, we stimulate three tactile receptors based on detection thresholds as described by Konyo et al. [7], i.e.: FA1: most sensitive between 25–40 Hz; FA2: most sensitive between 200–250 Hz; and SA1: most sensitive at approximately less than 5 Hz. The detection threshold for each tactile receptor is illustrated in Figure 1.

Stimuli Design

Our vibrotactile stimuli design concept and the correlation between its variables are illustrated in Figure 2-a. Each stimulus was designed by the superposition of three haptic vibrations of different frequency ranges, i.e. the constructive interference of three different frequency ranges. Six variables for the stimuli design are described in Table 1.

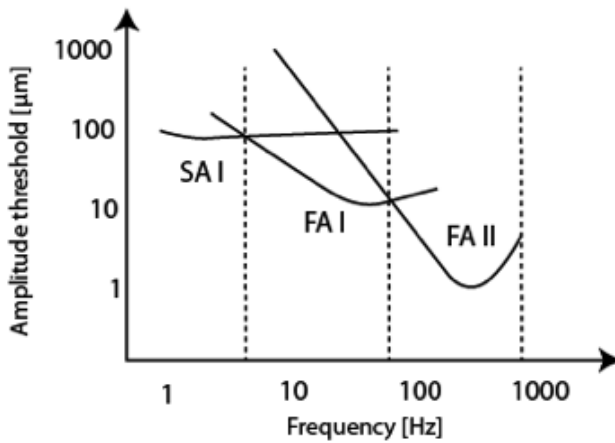


Fig. 1. Detection thresholds of vibratory stimuli based on Konyo et al. [7] that was originally based on Bolanowski et al. [8]

Table 1. Amplitude and frequency variables for a given force pattern

Amplitude Variables	Receptor Target	Amplitude Range	Frequency Variables	Receptor Target	Frequency Range (Hz)
Amplitude_FA1	FA1 (Meissner)	0 – 45	Frequency_FA1	FA1 (Meissner)	25– 40
Amplitude_FA2	FA2 (Pacini)	0 – 12	Frequency_FA2	FA2 (Pacini)	200 – 250
Amplitude_SA1	SA1 (Merkel)	0 – 60	Frequency_SA1	SA1 (Merkel)	0.4– 7

Stimuli Playback

100 stimuli were generated in this experiment and evaluated by our subject participants by Semantic Differential (SD) test. The values for six variables of haptic stimulus, as described in Table 1 were chosen randomly. There were no stimuli with identical combination of those six variables. In this experiment, vibrotactile stimuli were displayed using vibrotactile actuators (Figure 2-b).

4.3 Onomatopoeias

Japanese onomatopoeias were used for subjective evaluation test (Semantic Differential Test). This study adopts 42 onomatopoeias from Hayakawa et al. [6] and Japanese Onomatopoeia Dictionary [9]. After preliminary testing, 17 onomatopoeias were selected for Semantic Differential Test and constructing Principal Component Analysis (PCA) Diagram. List of onomatopoeias and its English equivalent is described in Table 2 [10]. However, the English equivalent in Table 2 may have different meaning, which depends on its context.

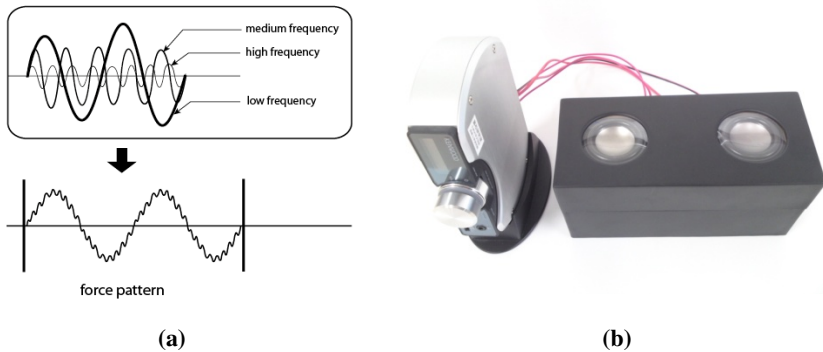


Fig. 2. (a) Stimuli design by superposition in this research (b) A prototype to display vibrotactile stimuli that consist of a pair of vibrotactile actuators and digital amplifier

4.4 Procedure

The experiment was conducted in a room with minimum noise and controlled temperature. The stimuli were generated by vibrotactile actuators as described in Figure 2-b. The stimuli were continuously played while the subject giving scores to Semantic Differential (SD) test. In this study, we used 7-point Likert scale SD questionnaire, both end of bipolar scale consists of “strongly felt” and “not felt at all.” There were 17 onomatopoeias and 100 stimuli for SD test, therefore we had 1700 set of data from each participant.

The experiment was held in an hour time limit, i.e. each participant was constrained not to answering more than one hour in a day. If participant couldn’t finish all 100 stimuli within one hour, the experiment was rescheduled on a latter day.

This experiment took approximately a month to finish. All of the data extracted in this experiment were analyzed using R statistics [11].

Table 2. List of onomatopoeias for subjective evaluation questionnaire

Onomatopoeia (in Japanese)	English Equivalent	Onomatopoeia (in Japanese)	English Equivalent
Kasakasa	A coarse, dried out feeling	Nurunuru	Greasing, soaping, making slippery
Gasagasa	A coarse, dried out feeling, coarser than kasakasa	Nechanecha	Adhesive, like glue, viscous, greasy
Kunyakunya	Soft, flexible, supple	Nechonecho	Slimy*
Gunyagunya	Soft, disfigured	Korikori	Scraping, crunchy
Gorigori	Hard, having a hard core	Jyarijyari	Sandy, gritty
Sarasara	Smooth, light, dry	Jyoriyori	Bristly*
Subesube	Smooth, slippery	Kochikochi	Sore and stiff
Tsurutsuru	Smooth, slippery (stronger than subesube)	Nyurunyuru	Squirm, slip away
Numenume	Smooth, slimy		

*) The English equivalent could not be found in the reference and a Japanese native speaker were asked to describe the meaning

5 Results

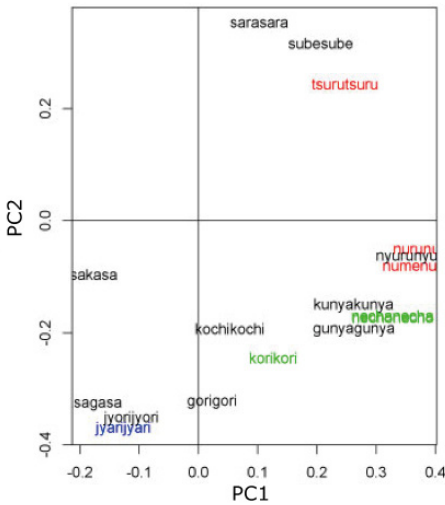
The result of Principal Component Analysis (PCA) of 17 adjectives is described in Table 3 and its visualization is depicted in Figure 3. In this study, 3 principal components extracted cumulative explained variance of 56.52%, which means that the current model can explain 56.52% of the overall touch perception.

The first principal component (PC1, 27.17% of variance) has positive correlation with “nurunuru” and “numenume”; Second principal component (PC2, 19.27% of variance) positively correlate with “sarasara” and “subesube,” and negatively correlate with “jyarijyari” and “jyoriyori”; Third principal component (PC3, 10.07% of variance) positively correlate with “kochikochi” and “korikori,” and negatively correlate to “kunyakunya” and “gunyagunya.”

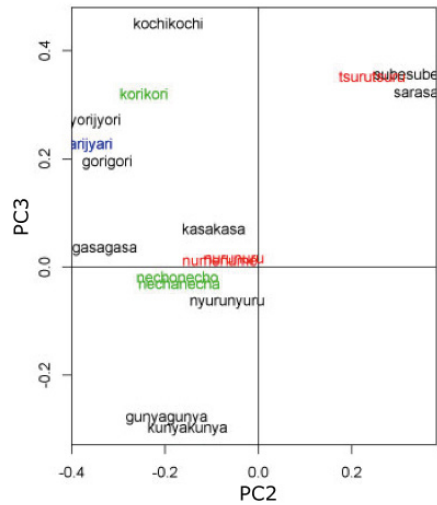
Two-dimensional plotting of PC1 & PC2, and PC2 & PC3 are depicted in Figure 3-a and Figure 3-b, respectively.

Table 3. Component loadings for 17 onomatopoeias for 3 Principal Components

Onomatopoeia	PC1	PC2	PC3	Onomatopoeia	PC1	PC2	PC3
Kasakasa	-0.19	-0.10	0.07	Nurunuru	0.38	-0.05	0.01
Gasagasa	-0.18	-0.33	0.03	Nechanecha	0.33	-0.17	-0.03
Kunyakunya	0.26	-0.15	-0.30	Nechonecho	0.33	-0.17	-0.02
Gunyagunya	0.26	-0.19	-0.28	Korikori	0.13	-0.25	0.32
Gorigori	0.02	-0.32	0.19	Jyarijyari	-0.13	-0.37	0.22
Sarasara	0.10	0.35	0.32	Jyoriyori	-0.11	-0.35	0.27
Subesube	0.21	0.32	0.36	Kochikochi	0.05	-0.19	0.45
Tsurutsuru	0.25	0.24	0.35	Nyurunyuru	0.36	-0.07	-0.07
Numenume	0.37	-0.08	0.01				



(a)



(b)

Fig. 3. Texture phase diagram of 17 onomatopoeias based on Principle Component Analysis

6 Discussion

6.1 Analysis of Texture Phase Diagram

Each principal component can be renamed according to the adjectives that have strong correlation with it. Another matter that needs to be considered is the

comparison to preceding researches by Hayakawa et al. [6] and Hollins et al. [4]. Our experiment result suggests that the adjectives that correlate with PC1, PC2, and PC3 represents moisture, friction, and hardness, respectively. This finding is in accordance with Hayakawa et al. (friction, hardness, and moisture) and similar with Hollins et al. who propose softness-hardness and roughness-smoothness as two of the most important element in tactile sensation.

The plotting of 3 basic components of artificial vibrotactile stimuli, i.e.: FA1 (~30 Hz), FA2 (~225 Hz), and SA1 (~5 Hz) are described in Figure 4. This plotting is approximate and based on average score of each adjective. FA1 corresponds to “jyarijyari,” “jyoriijyori,” and “gasagasa”; FA2 corresponds to “gasagasa” and “kasakasa”; and SA1 corresponds to “kunyakunya” and “gunyagunya.”

The superposition of more than one basic component (FA1, FA2, or SA1) results in a combined properties from both components in the new stimulus. For example, combining SA1 and FA1 will result in vibrotactile stimuli that correlate with “kunya-kunya,” “gunyagunya,” “jyariijyari,” “jyoriijyori,” and “gasagasa.” Furthermore, the non-existence of a basic component (FA1, FA2, or SA1) will cause a haptic stimulus to be located on the opposite polar of its counterpart. For example, stimulus with very low Amplitude_FA1 is located nearby “sarasara,” “subesube,” and “tsurutsuru.”

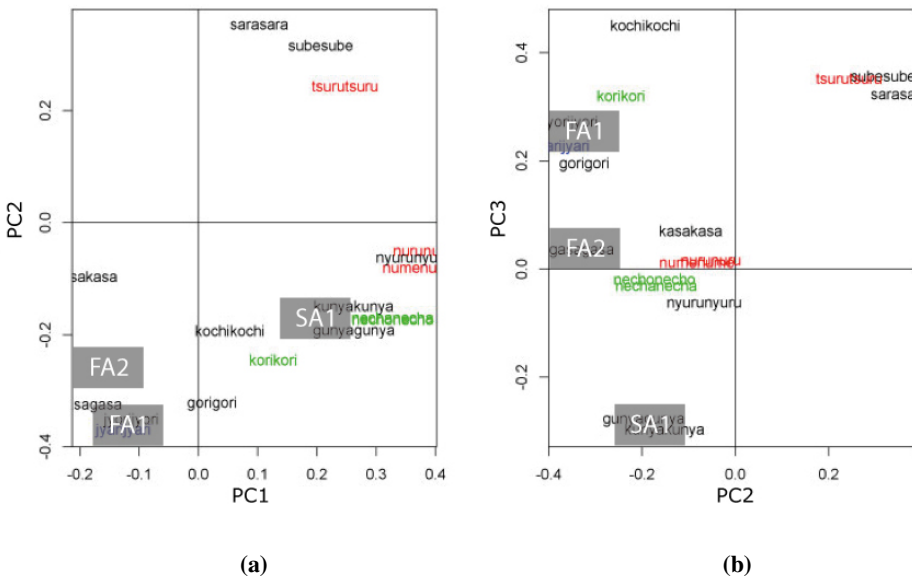


Fig. 4. Texture phase diagram of 17 onomatopoeias based on Principle Component Analysis and artificial stimuli plot

6.2 Comparison between Texture Phase Diagram of Physical Materials and Artificial Stimuli

This study gives similar result to previous studies in texture phase mapping, as described in section 6.1. However, there are some differences that are discovered in this study. Although our experiments yield the same principal components as Hayakawa et al., we found that the order of principal components is different. In texture phase diagram of physical materials by Hayakawa et al., the order of principle components are friction, hardness, and moisture, as measured by the number of variance explained. However, we found that dampness (PC1, 27.17% of variance) is the principle component that explains most of the variance, followed by friction (PC2, 19.27% of variance) and hardness (PC3, 10.07% of variance). We argue that vibrotactile actuators cannot adequately generate hardness sensations and more suitable to be used for generating friction or moisture sensations. Directly stimulating Meissner or Pacinian corpuscle by vibrotactile actuators can emulate those sensations.

7 Conclusion and Future Works

This study has proposed a new classification method of tactile sensations generated by artificial vibrotactile stimuli. A texture phase diagram has been developed that can be used to explain the correlation between artificial vibrotactile stimuli and tactile perception. In addition, the comparison to tactile sensations generated by physical materials has also been discussed.

Furthermore, although the explained cumulative variance is relatively low (56.52%), it suggests that artificial vibrotactile stimuli may not adequately emulate tactile sensations that are generated by genuine physical materials. However, this result proposes a new insight towards possible applications of artificial tactile stimuli in the future.

Future works will include expanding the model with possible addition of new onomatopoeias/adjectives, developing preliminary device prototype that utilized touch perception model in this research, and developing its corresponding user interface. Specifically, the future works will be aimed at developing comprehensive tactile mapping and symbol that can be used to convey message universally.

Acknowledgement. This work was supported by JSPS KAKENHI Grant Number 245436.

References

1. Chen, X., Shao, F., Barnes, C., Childs, T., Henson, B.: Exploring relationships between touch perception and surface physical properties. *International Journal of Design* 3(2), 67–77 (2009)
2. Shirado, H., Maeno, T.: Modeling of Human Texture Perception for Tactile Displays and Sensors. In: *The First Joint Eurohaptics Conference and Symposium on Haptic Interface for Virtual Environment and Teleoperator Systems*, Pisa, Italy, pp. 629–630 (2005)

3. Shirado, H., Konyo, M., Maeno, T.: Modeling of Tactile Texture Recognition Mechanism. *Journal of the Japan Society of Mechanical Engineers* 73(733), 2514–2521 (2007) (in Japanese)
4. Hollins, M., Faldowski, R., Rao, S., Young, F.: Perceptual dimensions of tactile surface texture: A multidimensional scaling analysis. *Perception & Psychophysics* 54, 697–705 (1993)
5. Tiest, W.M.B., Kappers, A.M.L.: Analysis of haptic perception of materials by multidimensional scaling and physical measurements of roughness and compressibility. *Acta Psychologica* 121, 1–20 (2006)
6. Hayakawa, T., Matsui, S., Watanabe, J.: Classification method of tactile textures using onomatopoeias. *Journal of The Virtual Reality Society of Japan* 15(3), 487–490 (2010) (in Japanese)
7. Konyo, M., Yoshida, A., Tadokoro, S., Saiwaki, N.: A tactile synthesis method using multiple frequency vibrations for representing virtual touch. In: *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, Alberta, Canada, pp. 1121–1127 (August 2005)
8. Bolanowski, S.J., Gescheider, G.A., Verillo, R.T., Checkosky, C.M.: Four channels mediate the mechanical aspects of touch. *Journal of the Acoustical Society of America* 84, 1680–1694 (1988)
9. Ono, M.: *Giongo, Gitaigo 4500: Nihongo Onomatope Jiten (4500 Onomatopoeias and Sound-symbolic Words: Japanese Dictionary of Onomatopoeias)*. Shogakkan (2007) (in Japanese)
10. Kamermans, M.: <http://www.nihongoresources.com/>
11. R Development Core Team, R: *A Language and Environment for Statistical Computing* (2010)

Preliminary Design of Haptic Icons from Users

Wonil Hwang and Dongsoo Kim

Department of Industrial and Information Systems Engineering, Soongsil University, Korea
{wonil, dskim}@ssu.ac.kr

Abstract. Haptic icons are useful for blind people as well as normal people to perceive information from their environments. Thus, lots of efforts were given to designing usable haptic icons, but not much progress was made in designing haptic icons so far, in terms of variety and intuitiveness. The purpose of this study is to investigate how to match vibrotactile stimuli with representational information or abstract concepts to design a variety of and intuitive haptic icons. We employed the bi-directional approach to ask users about their association between representational information/abstract concepts and perceived vibrotactile stimuli. Two-staged experiments were conducted with forty participants. From the experiments, verbal descriptions corresponding to each of 36 vibrotactile stimuli and drawings of vibration corresponding to each of 27 representational information/abstract concepts in the context of human-computer interaction were collected. We can conclude that the associations that users described from these experiments would provide the foundation for designing more intuitive haptic icons in enough variety.

Keywords: Haptic icons, vibrotactile stimuli, representative information, abstract concepts, intuitiveness.

1 Introduction

Haptics comprises the studies related to delivering information based on senses of touch and to developing instruments to facilitate information delivery. Although visual and auditory channels have been considered as two main communication channels of human bodies, recently, the haptic channel based on vibrotactile stimuli has been actively studied to support or replace visual and auditory channels [6, 8]. For example, haptic navigation aids for automotive drivers and surgical instruments based on haptic feedback are haptic applications as communication channels. And also, especially, the haptic devices as communication media help blind peoples to easily recognize information from their environments [3]. Thus, haptic channels are useful for both of normal people and blind people to perceive information from their environments.

Likewise visual and auditory displays, the haptic displays are utilized as a medium between users and information providers for effective and efficient communication. Visual and auditory displays may often employ visual and auditory icons for users to

intuitively understand information [1]. In the same vein, haptic displays need haptic icons that are as much usable as visual and auditory icons. Even though lots of efforts were given to designing usable haptic icons [2, 4, 7], not much progress was made in designing haptic icons so far, in terms of variety and intuitiveness. Therefore, this study was motivated on how we design haptic icons that have intuitiveness and intimacy with users.

The purpose of this study is to investigate how to match vibrotactile stimuli with representational information or abstract concepts to design a variety of and intuitive haptic icons. Haptic icons should be distinguishable from each other and easily learned by users. So, this study tries to find what kinds of representational information or abstract concepts users associate with specific vibrotactile stimuli.

2 Methods

We employed the bi-directional approach to ask users about their association between representational information/abstract concepts and perceived vibrotactile stimuli: 1) asking users what kinds of representational information/abstract concepts are to be delivered by each vibrotactile stimulus; and 2) asking users to draw proper vibrotactile stimulus to deliver each representational information/abstract concepts. By this bi-directional approach, two-staged experiments were conducted and a total of forty participants took part in the experiments.

2.1 Apparatus

In order to generate vibrotactile stimuli that were used in the first stage, we utilized the vibration excitation system, which was designed to generate sinusoidal vibration with a range of frequency and amplitude. This vibration excitation system included the mini-shaker (Brüel and Kjær Type 4810) and a programmable function generator (Tabor Electronics WW5062) as the key components. The operating frequency range of the mini-shaker is DC - 18,000 Hz, and the maximum bare table acceleration amounts to 550 m/s² (55.1 g) in the frequency range of 65 Hz - 4,000 Hz. The programmable function generator was controlled by a waveform creation software (ArbConnection 4.x) in PC.

2.2 Experiments

In the first stage, 20 participants (8 males and 12 females; mean: 23.9 years old and standard deviation: 1.65 years old) took part in the experiment. They were presented a total of 36 distinguishable vibrotactile stimuli, which were generated by two kinds of combination: (1) Five types of frequency (320, 160, 80, 40, 20Hz) and four types of rhythm (vibrating in 100%, 75%, 50%, 25% of a second) with 20dB sinusoidal vibration; and (2) Five types of macro-wave pattern (horizontal, increasing, decreasing, increasing-decreasing, decreasing-increasing patterns) and four types of rhythm (vibrating in 100%, 75%, 50%, 25% of a second) with 160Hz and 20dB

sinusoidal vibration. And considering human-computer interaction context they were asked to describe a proper representational information/abstract concept that was to be delivered by each vibrotactile stimulus after perceiving each of 36 distinguishable vibrotactile stimuli in random order.

In the second stage, 27 representational information/abstract concepts were selected by the results of the first stage experiments and other 20 participants (10 males and 10 females; mean: 23.9 years old and standard deviation: 1.62 years old) participated in the experiments. They were asked to freely draw a proper vibrotactile stimulus with frequency, rhythm and macro-wave pattern that was to deliver each of 27 representational information/abstract concepts in random order.

3 Results

The data resulted from the two-staged experiments were quite qualitative, because we collected the verbal descriptions corresponding to each of 36 vibrotactile stimuli from the first stage experiments and drawings of vibration corresponding to each of 27 representational information/abstract concepts in the context of human-computer interaction from the second stage experiments. Despite such qualitative data we tried to analyze them in a way similar to quantitative analysis by counting the frequency of verbal descriptions and coding the drawings with 3-digit numbers.

3.1 First Stage Experiments

From the first stage experiments, 20 participants represented each of 36 vibrotactile (or haptic) stimuli as 15 to 28 descriptions that were used in the context of human-computer interaction. Because many of these descriptions were duplicated one another in their meaning, we grouped them and selected most frequent 27 descriptions as representative information/abstract concepts for vibrotactile (or haptic) stimuli (see the first column of Table1).

3.2 Second Stage Experiments

From the second stage experiments, 20 participants drew a vibration for each of 27 representative information/abstract concepts, and we coded them with 3-digit numbers (see Figure 1 and 2) as follows: (1) First digit: frequency type (1: high, 2: middle, 3: low); (2) second digit: macro wave type (1: horizontal, 2: increasing, 3: decreasing, 4: increasing & decreasing, 5: decreasing & increasing, 6: stepping down, 7: stepping up); and (3) third digit: rhythm type (1: vibrating in 100%, 2: vibrating in 75%, 3: vibration in 50%, 4: vibrating in 25% of a second). And then we made a frequency table per representative information/abstract concept, and resulted in

recommended haptic stimuli for each of representative information/abstract concept based on the highest frequency (see the second column in Table 1). We calculated Herfindahl-Hirschman Index (HHI), which indicated the concentration ratio (originally, it measures the amount of competition in the market) [5], and used it as a measure of consensus among 20 participants about haptic stimuli for each of representative information/abstract concept (see the last two columns in Table 1).

As shown in Table 1, 15 distinguishable haptic stimuli are recommended to represent specific information/abstract concepts, even though some of recommended haptic stimuli repeatedly appear for representative information/abstract concepts. For example, '211', which is coded for vibration with middle frequency, horizontal macro wave type and vibrating in 100% of a second, appears for 8 representative information/abstract concepts. If we apply 'normalized HHI > 0.025' to find relatively good consensus among 20 participants about haptic stimuli for each of representative information/abstract concept, we can find that there are good consensus for 12 representative information/abstract concepts, such as clicks/selecting, dragging/moving, detecting virus, program stopped/no response, error, completing actions, inputting data/typing, double-clicks/executing, in the middle of action/buffering/installing, deleting, alarm/message, booting.

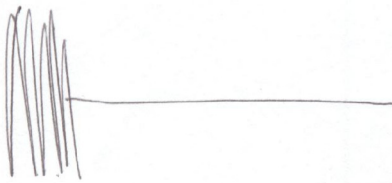


Fig. 1. An example of drawing coded as '214' and representing 'clicks/selecting'

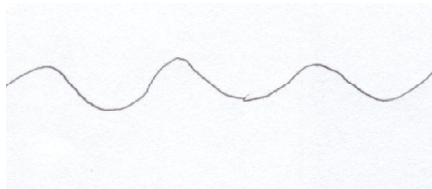


Fig. 2. An example of drawing coded as '311' and representing 'dragging/moving'

Table 1. Drawing results of haptic stimuli

Information delivered by haptic stimuli	Recommended haptic stimuli ¹	Percent of votes for a recommended haptic stimulus	HHI	Normalized HHI
Clicks/Selecting	214	45.0%	0.2600	0.1675
Dragging/Moving	311	28.6%	0.1338	0.0616
Detecting virus	211	22.7%	0.1446	0.0496
Program stopped/No response	214	25.0%	0.1150	0.0469
Error	211, 213	20.0%	0.1150	0.0413
Completing actions	211	22.7%	0.0992	0.0391
Inputting data/Typing	213	20.0%	0.1150	0.0345
Double-clicks/Executing	213	20.0%	0.1150	0.0345
In the middle of action/Buffering/Installing	211, 311	18.2%	0.1033	0.0343
Deleting	133	20.0%	0.1200	0.0320
Alarm/Message	213	20.0%	0.1050	0.0304
Booting	213	20.0%	0.0950	0.0254
New windows/Pop-ups	214	20.0%	0.0900	0.0250
Sending-receiving files/Up-loading/Down-loading	211	20.0%	0.0850	0.0240
Closing programs or windows	133, 211, 214	13.6%	0.0868	0.0215
Locking up	114	18.2%	0.0785	0.0209
Scrolling down	214, 311	15.0%	0.0900	0.0200
Undoing	114, 134	15.0%	0.0850	0.0196
Warning	111, 212	15.0%	0.0950	0.0196
Computer shut-down	211	15.0%	0.0800	0.0143
Creating folders, files or shortcuts	221	15.0%	0.0800	0.0143

¹ Three-digit means 'frequency type - macro wave type - rhythm type': (1) Frequency type (1: high, 2: middle, 3: low); (2) macro wave type (1: horizontal, 2: increasing, 3: decreasing, 4: increasing & decreasing, 5: decreasing & increasing, 6: stepping down, 7: stepping up); (3) rhythm type (1: vibrating in 100%, 2: vibrating in 75%, 3: vibration in 50%, 4: vibrating in 25% of a second).

Table 1. (Continued)

Information delivered by haptic stimuli	Recommended haptic stimuli	Percent of votes for a recommended haptic stimulus	HHI	Normalized HHI
Switching windows	214	15.0%	0.0800	0.0143
Minimizing windows	331	15.0%	0.0750	0.0133
Printing	Six stimuli	10.0%	0.0800	0.0092
Refreshing	Four stimuli	9.1%	0.0620	0.0068
Connecting internet or messenger	211, 212, 222	10.0%	0.0650	0.0066
Maximizing windows	124, 223, 321	10.0%	0.0650	0.0066

4 Conclusions and Discussions

From the two-staged experiments, we found what kinds of association between representational information/abstract concepts and vibrotactile stimuli users had in mind. First, from the first stage experiments we could select a total of 27 representative information/abstract concepts related to human-computer interaction from 36 distinguishable vibrotactile (or haptic) stimuli. It means that these 27 concepts are good candidates for information that can be delivered by haptic stimuli. Second, from the second stage experiments we could find which haptic stimuli could show relatively good representation for each of 27 concepts. Third, by introducing Herfindahl-Hirschman Index as a measure of consensus we could show how much users agreed that each of representative information /abstract concepts could be delivered by specific haptic stimuli. For example, ‘clicks/selecting’ is the representative information/abstract concept that users give the biggest consensus. Even though this study has some limitations as a preliminary study, including small sample size, the associations that users described from these experiments would provide the foundation for designing more intuitive haptic icons in enough variety.

Acknowledgements. This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2011-0013713).

References

1. Blattner, M.M., Sumikawa, D.A., Greenberg, R.M.: Earcons and icons: Their structure and common design principles. *Human-Computer Interaction* 4, 11–44 (1989)

2. Brewster, S.A., Brown, L.M.: Tactons: Structured tactile messages for non-visual information display. In: Proc. of the 5th Australasian User Interface Conference, Sydney, Australia, pp. 15–23 (2004)
3. Colwell, C., Petrie, H., Kornbrot, D., Hardwick, A., Furner, S.: Haptic virtual reality for blind computer users. In: Proc. of the 3rd International ACM Conference on Assistive Technologies, New York, NY, USA, pp. 92–99 (1998)
4. Enriquez, M., MacLean, K.E., Chita, C.: Haptic phonemes: Basic building blocks of haptic communication. In: Proc. of the 8th International Conference on Multimodal Interfaces, Los Alamitos, CA, USA, pp. 302–309 (2006)
5. Hirschman, A.O.: The paternity of an index. *The American Economic Review* 54(5), 761–762 (1964)
6. Hwang, J., Hwang, W.: Vibration perception and excitatory direction for haptic devices. *Journal of Intelligent Manufacturing* 22, 17–27 (2011)
7. Hwang, J., Hwang, W.: Generation of effective vibrotactile stimuli to convey information. *ICIC Express Letters* 7(5), 1637–1641 (2013)
8. Ji, Y.G., Lee, K., Hwang, W.: Haptic perceptions in the vehicle seat. *Human Factors and Ergonomics in Manufacturing & Service Industries* 21(3), 305–325 (2011)

Assessing the Effectiveness of Vibrotactile Feedback on a 2D Navigation Task

Wooram Jeon, Yueqing Li, Sangwoo Bahn, and Chang S. Nam

Edwards P. Fitts Department of Industrial and Systems Engineering,
North Carolina State University, Raleigh, NC, 27695-7906, USA
{wjjeon, yli48, sbahn, csnam}@ncsu.edu

Abstract. The effect of vibrotactile parameters were investigated on a 2D navigation task. Participants performed a simple navigation task reproducing directional information presented by a series of vibrotactile stimuli consisting of different levels of amplitude and frequency. Task completion time and degree of annoyance were measured. The results demonstrated that both frequency and amplitude had a significant effect on the responses. In addition, interaction effects between the two parameters were found on the responses. It was concluded that user performance and comfort are significantly affected by frequency and amplitude. The results give some insight into designing navigating information presented by vibrotactile display for visually impaired people. More studies with people with visual impairment and manipulation of other vibrotactile parameters are recommended to be applicable to the potential research.

Keywords: Tactile display, vibrotactile, haptic, navigation.

1 Introduction

Most traditional computer-based machines have relied on visual presentation to deliver information to users. However, there are cases in which visual displays are inappropriate. For example, when interacting with in-vehicle systems, the visual sensory channel can be pressured from the constant information in the traffic scene itself. This can result in the cognitive capacities that drivers have at their disposal being overloaded (Van Erp & Van Veen, 2004). Another possible case in which non-visual communication is required is for people with a visual impairment. Therefore, it is important to consider using alternative modalities through which information can be presented. The use of the auditory sense as an alternative channel for communication has been widely investigated (Blattner, Sumikawa, & Greenberg, 1989). An alternative modality which might be beneficial in these cases is the sense of touch.

The sense of touch, referred to as haptics, has been used to aid communication for people with visual impairments. For example, one use of the sense of touch by people with visual impairment is the Braille system, which enables them to read text. Another method with potential for communication via haptics is vibration. It was proposed that vibrotactile stimuli could be used to present information by manipulating different parameters of vibration (Geldard, 1960).

Vibrotactile displays have recently been common in a variety of devices, such as mobile phones, handheld PCs and game console controllers. However, the types of vibration used in such devices are generally very simple and does not provide much information as a means of communication. Researchers have recently begun to explore the possibility of implementing vibrotactile displays for presenting more complex information, such as in navigation (Van Erp & Van Veen, 2001; Van Erp & Van Veen, 2004). However, most research uses vibrotactile displays/feedback as secondary information channel in a specific application and only tests whether it is applicable and whether the user performance is enhanced. Little research has been conducted to investigate the effect of vibrotactile feedback as primary modality information on enhancing users' spatial orientation.

Vibrotactile navigation systems have been developed and investigated by many studies to aid navigation for a variety of fields including car drivers, pilots, and people with visual impairment. Van Erp and van Veen (2001) designed vibrotactile icons for an in-vehicle navigation system, using vibrotactile devices mounted in a car seat. In addition, Van Veen and Van Erp (2000) have investigated the use of a vibrotactile vest to provide navigation information for airplane pilots. It was found that tactile display would be particularly useful when pilots are in harsh conditions, and tactile information might be more readily received. In addition to aiding navigation in vehicles, vibrotactile feedback has also been used to help blind or visually impaired people to navigate. For visually impaired people, Ross and Blasch (2000) designed a wearable tactile display, which indicated whether the user was walking in the right direction or if a change of direction was needed (Ghiani, Leporini, & Paternò, 2008).

Ghiani et al. (2008) developed mobile museum guide which provides vibrotactile feedback for blind users. It was designed to be easily plugged into PDAs to assist blind users in orientation. Geldard (1960) proposed that the main parameters of vibration are intensity (amplitude), frequency, signal (waveform) duration, rhythm, and spatial location. This study investigated intensity and frequency as vibrotactile parameters for communicating navigation information. Intensity refers to the square of the amplitude of the signal. Since the terms intensity and amplitude are often used interchangeably, the term amplitude is mainly used in this study. Frequency refers to the rate of vibration and is expressed in Hertz (Hz). Simple navigation tasks were proposed with manipulation of these two parameters. We were interested in examining how the vibrotactile parameters affected user performance and annoyance on 2D navigation task.

2 Methodology

2.1 Participants

Twelve participants were recruited from the student population at a local University. Their ages ranged from 19 to 23 years ($M = 21.9$, $SD = 2.4$). None of subjects had prior experience with a similar type of the experiment.

2.2 Apparatus

A computer-controlled system was designed to generate vibrotactile information on a vibrotactile array. The computer controlled the trial conditions, manipulated parameters, and logged response data. The tactile display consisted of an elastic belt that subjects wore around their abdomen, on which the vibrotactile tactor array was attached. The tactor array was composed of four (C2 tactors (Engineering Acoustics, Inc). The C2 tactor incorporates a moving contactor that is lightly preloaded against the skin. When an electrical signal is applied, the contactor oscillates perpendicular to the skin, while the surrounding skin area is shielded with a passive housing. Four tactors contacted the left, right, front, and back of abdomen, to represent four directions (left(L), right(R), up(U), and down(D)) in the navigation task.

2.3 Task Design

The proposed navigation task was performed with the vibrotactile array belt which conveys direction information. The vibrotactile array belt has four tactors, each producing an independent vibration. A navigation task has a series of four vibrotactile stimulus. After the sequence of stimuli was presented, participants were required to mark a path in a grid paper with pencil. Each stimulus would represent one movement in the grid paper with specific direction. For example, if the sequence of vibrotactile stimulus is D-L-U-R and the participant correctly perceived the stimuli, a participant should draw a navigation path which is Down (D) first, Left (L), Up (U), and finally Right (R) from the starting point. Participants were asked to draw the path as quickly as possible and let the experimenter know when done with drawing.

The proposed navigation task needs participants' memory recall. The responses may be affected by a limited working memory capability. To this end, a screening test was designed to measure the participant's vibrotactile working memory capability. In the screening test, the number of stimulus in each sequence increased from 2 to 6. Based on the results from the screening task, participants with vibrotactile working memory less than 4 were removed from the data analysis.

2.4 Experiment Design and Variables

The experiment followed a factorial combination within-subject design. Amplitude and frequency were the independent variables. There were two levels of amplitude in the study. The amplitude level was created by choosing the gain value provided by the tactor controller: large amplitude ($A_1 = 4.1$), and small amplitude ($A_2 = 1.0$). Since only sinusoidal stimuli were available, frequency was the number of cycles of sinusoidal stimuli occurring in one second. There were three levels of frequency in the study: high frequency ($F_1 = 349$ Hz) (highest frequency available provided by the controller), medium frequency ($F_2 = 200$ Hz), and low frequency ($F_3 = 50$ Hz). Task completion time and the degree of user's annoyance were measured in terms of millisecond and rating scale 1 (small) though 7 (large) respectively.

2.5 Procedure

The participants received instructions describing the experiment. They were asked to read and sign an informed consent. After the belt was fastened, they were asked whether they could distinguish vibration with different levels of parameters. The experiment started with a screening test. The screening test included 15 trials with the number of vibrotactile stimuli varying from 2 to 6. Five minutes break was provided after the screening test. The experiment session was composed of 36 trials, with 4 vibrotactile feedback in each trial. Rest was provided every 10 trials and whenever requested by the participant. In both the screening test and the following experiment session, the trials were completely randomized to avoid the sequence effect. During the second half of the trials, the participants were asked to mark a degree of annoyance between 1 and 7 for each trial. The whole experiment lasted around 45 minutes.

3 Result

Each dependent variable was analyzed using an ANOVA with amplitude (high, low) x frequency (high, medium, low) factors. Post hoc analyses were completed using the Tukey HSD test with the alpha level set at .05. Table 1 summarizes the significant effects for the responses.

Table 1. Significant effects for performance parameters

Parameter	Effect	F-Value	p-value
Completion Time	Frequency	$F_{2,426} = 99.15$	<0.0001
	Amplitude	$F_{1,426} = 45.72$	<0.0001
	Frequency*Amplitude	$F_{2,426} = 14.12$	<0.0001
Annoyance	Frequency	$F_{2,426} = 33.03$	<0.0001
	Amplitude	$F_{1,426} = 31.55$	<0.0001
	Frequency*Amplitude	$F_{2,426} = 9.20$	0.0001

Result showed that frequency had a significant effect on task completion time ($F_{2,426} = 99.15$, $p = <0.0001$). Post analysis of frequency showed that participants had significantly shorter completion time in medium frequency ($M = 7686$, $SD = 4972$) than high frequency ($M = 10410$, $SD = 10239$), which also had significantly shorter completion time than low frequency ($M = 23625$, $SD = 15478$). Amplitude ($F_{1,426} = 45.72$, $p < 0.0001$) was also found to have significant effect. Post analysis of amplitude showed that participants had significantly shorter completion time in large amplitude ($M = 10565$, $SD = 10426$) than small amplitude ($M = 17250$, $SD = 14551$).

A significant interaction effect between frequency and amplitude was found, ($F_{2,426} = 14.12$, $p < 0.0001$). As Fig. 1 showed, the task completion time decreased as amplitude increased at all frequency levels. But, the decrease rate was much greater at low frequency than medium and high frequency.

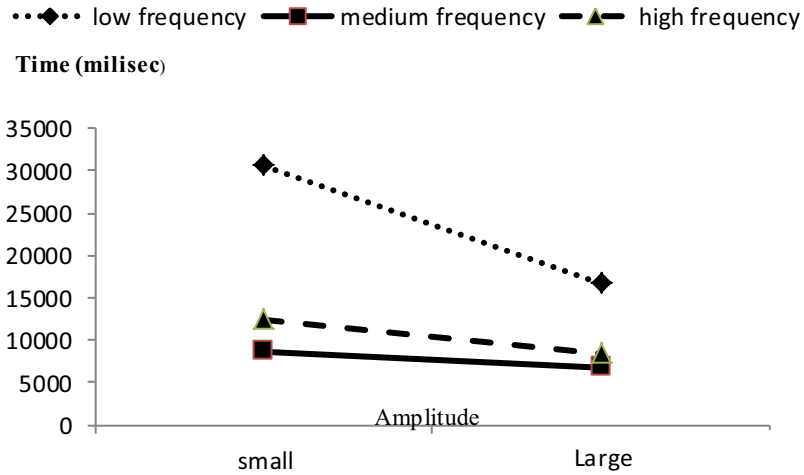


Fig. 1. Interaction effect between frequency and amplitude on task completion time

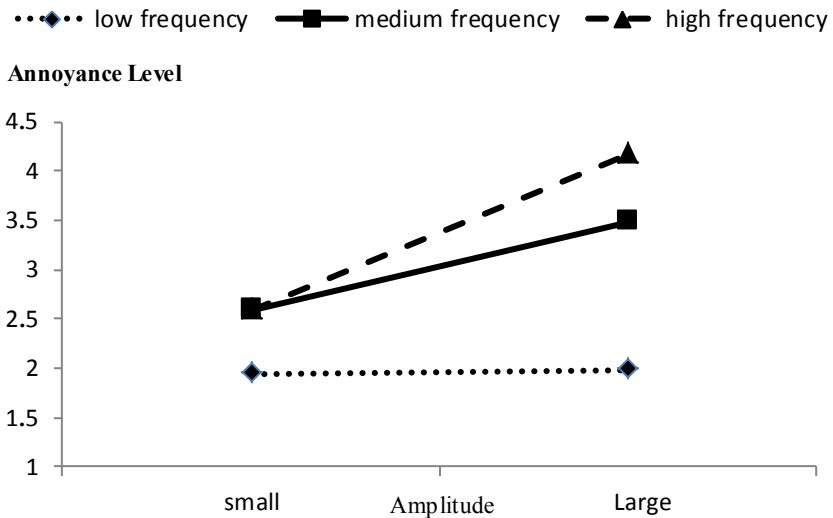


Fig. 2. Interaction effect between frequency and amplitude on annoyance

In addition, frequency had a significant effect on the annoyance level ($F_{2, 426} = 33.03, p < 0.0001$). Post analysis of frequency showed that participants had a significantly lower annoyance level in low frequency ($M = 1.96, SD = 1.65$) than medium ($M = 3.03, SD = 1.54$) and high frequency ($M = 3.38, SD = 1.68$). Amplitude ($F_{1, 426} = 31.55, p < 0.0001$) was also found to have significant effect. Post analysis of amplitude showed that participants had a significantly lower annoyance level in small amplitude ($M = 2.37, SD = 1.67$) than in the large amplitude ($M = 3.20, SD = 1.69$).

A significant interaction effect between frequency and amplitude was found ($F_{2, 426} = 9.20$, $p = 0.0001$). As Fig. 2 showed, the user annoyance level increased as amplitude increased at all frequency levels. However, the increase rate also increased as frequency level increased. In other words, the increase rate was largest at high frequency, and smallest at low frequency.

4 Discussion

The results of the study illustrated the effect of the parameters of the vibrotactile display in a navigation task. A significant effect of the two parameters on task completion time and user annoyance was found.

Amplitude is the most direct measure of whether the stimulus is strong or weak. Therefore, it is self-evident that vibrotactile stimuli with larger amplitude are easier for participants to perceive than those with smaller amplitude. In terms of annoyance, vibrators can generate sufficient heat to cause a painful sensation of heat on the user's skin. In addition, tactile stimuli are hard to ignore if the user does not want to sense them (Van Erp & Van Veen, 2001). However, in the study, participants felt moderately annoyed in the large amplitude condition. Since the on-time duration of the tactile display was short (shorter than 2 seconds) users may not have been annoyed regardless of amplitude levels. To better understand how the participants' annoyance level changes, a time series analysis of annoyance level may help. To improve the task performance without sacrificing user friendliness, more levels of amplitude should be tested in future research.

5 Conclusion

This study investigated the effect of vibrotactile feedback on a navigation task. Vibrotactile amplitude and frequency were manipulated to present different patterns of vibrotactile stimuli. The participants performed 2D navigation tasks with vibrotactile stimuli conveying the moving direction. The results showed that all the parameters have a significant effect on user performance (task completion time) and comfort (the degree of annoyance). These results should provide insight to the real-world applicability of the vibrotactile feedback as a primary modality information provider.

In the present study all subjects were healthy people without any visual impairment. It is expected that vibrotactile would be a significant aid to those with visual impairment. Therefore, more studies with the participants having visual impairment are needed to ensure that the results of the study will be applicable to those potential users. In addition, other vibrotactile parameters, such as on-time duration or rhythm, should be investigated.

References

1. Alahakone, A., Senanayake, S.A.: Vibrotactile feedback systems: Current trends in rehabilitation, sports and information display. Paper Presented at IEEE/ASME International Conference on the Advanced Intelligent Mechatronics, AIM 2009, pp. 1148–1153 (2009)
2. Blattner, M.M., Sumikawa, D.A., Greenberg, R.M.: Earcons and icons: Their structure and common design principles. *Human-Computer Interaction* 4(1), 11–44 (1989)
3. Cholewiak, R.W., Collins, A.A.: The generation of vibrotactile patterns on a linear array: Influences of body site, time, and presentation mode. *Attention, Perception, & Psychophysics* 62(6), 1220–1235 (2000)
4. Geldard, F.A.: Some neglected possibilities of communication. *Science* (1960)
5. Gemperle, F., Ota, N., Siewiorek, D.: Design of a wearable tactile display. Paper Presented at Proceedings of the Fifth International Symposium on the Wearable Computers, pp. 5–12 (2001)
6. Ghiani, G., Leporini, B., Paternò, F.: Vibrotactile feedback as an orientation aid for blind users of mobile guides. Paper presented at the Proceedings of the 10th International Conference on Human Computer Interaction with Mobile Devices and Services, pp. 431–434 (2008)
7. Rothenberg, M., Verrillo, R.T., Zahorian, S.A., Brachman, M.L., Bolanowski Jr, S.J.: Vibrotactile frequency for encoding a speech parameter. *The Journal of the Acoustical Society of America* 62, 1003 (1977)
8. Salzer, Y., Oron-Gilad, T., Ronen, A.: Vibrotactor-belt on the thigh - directions in the vertical plane. In: Kappers, A.M.L., van Erp, J.B.F., Bergmann Tiest, W.M., van der Helm, F.C.T. (eds.) *EuroHaptics 2010, Part II. LNCS*, vol. 6192, pp. 359–364. Springer, Heidelberg (2010)
9. Schiff, W., Foulke, E. (eds.): *Tactual perception: a sourcebook*. Cambridge University Press (1982)
10. Van Erp, J.B., Van Veen, H.: Vibro-tactile information presentation in automobiles. Paper presented at the Proceedings of Eurohaptics, pp. 99–104 (2001)
11. Van Erp, J.B., Van Veen, H.A.: Vibrotactile in-vehicle navigation system. *Transportation Research Part F: Traffic Psychology and Behaviour* 7(4), 247–256 (2004)
12. Van Erp, J.: Presenting directions with a vibrotactile torso display. *Ergonomics* 48(3), 302–313 (2005), doi:10.1080/0014013042000327670

Magnetic Field Based Near Surface Haptic and Pointing Interface

Kasun Karunanayaka, Sanath Siriwardana, Chamari Edirisinghe,
Ryohei Nakatsu, and Ponnampalam Gopalakrishnakone

KEIO NUS Cute Center, Interactive and Digital Media Institute,
National University of Singapore, Singapore
{g0800474, sanath, chamari, elenr, antgopal}@nus.edu.sg

Abstract. Magnetic field based Near Surface Haptic and Pointing Interface is a new type of pointing interface which provides mouse interactions, haptic feedback and other enhanced features. It could also be configured as a haptic display, where users can feel the basic geometrical shapes in the GUI by moving the finger on top of the device surface. These functionalities are attained by tracking 3D position of a neodymium magnet, using Hall Effect sensors grid and generating like polarity haptic feedback using an electromagnet array.

Keywords: Pointing interface, haptic mouse, near surface haptic feedback, tactile display.

1 Introduction

The pointing devices are widely used as input interfaces to control and provide data to the graphical user interfaces (GUI) using physical gestures [1]. Movements and commands send by those pointing devices are echoed on the screen by movements of the mouse pointer (or cursor) and other visual changes. Mouse is the most common pointer device used today and there are also other devices such as track pad, track ball, stylus, and joystick. Recently, there were some attempts to add haptic feedback sensations to the pointing input interfaces. Most of those implementations used technologies such as piezoelectric actuators, pneumatic actuators and vibration motors. It can be understood that the addition of haptic sensations could enhance the attachment between the user and the computer [10].

Haptic Mouse is a new type of pointing interface which provides mouse interactions, haptic feedback and other enhanced features. The key advantage of this system over the other haptic pointer interfaces is that users do not required to touch the surface of the device. Instead the users could move the neodymium magnet worn on the fingertip near to the device surface and controls the cursor movement (Fig.1). As a result, it enables the haptic sensations in 3D space which will be a novel experience. Cursor movements are handled once the north pole of the neodymium magnet face downwards and mouse commands can be execute when south pole of the neodymium magnet downwards.

Different haptic sensations provided by this system can be felt like attraction, repulsion and various patterns of vibrations. Furthermore, this interface can be configured as a Haptic display. It is possible for a user to move his/her finger on top of the surface and sense the basic shapes of the objects on the screen. Simple geometrical shapes which are bigger than 200 pixels can be sensed and identified.

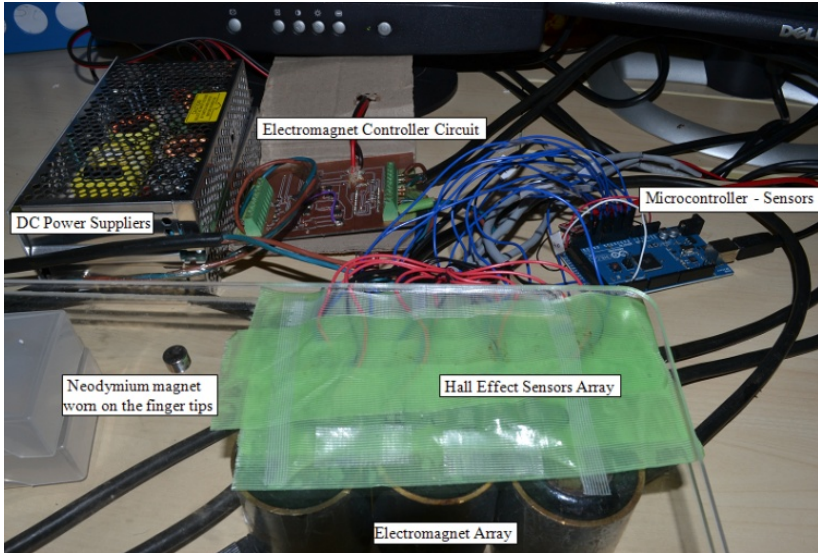


Fig. 1. Magnetic field based Near Surface Haptic and Pointing Interface system: User can move the neodymium magnet worn on the finger tip above the device surface and interact with the computer & sense haptic feedback for their inputs

2 Related Works

This section will discuss prior research with which the authors are arguing for the novelty value of the Magnetic field based Near Surface Haptic and Pointing Interface.

Liquid Interface [2] was a previous work of authors, which utilizes ferrofluid as an output display and input buttons embodied with musical notes. Using a matrix of Hall Effect sensors, magnetic fields generated by neodymium magnets worn on the fingertips are measured and then converted into signals that provide input capability. This input actuates an array of electromagnets and generates ferrofluid bubbles. By matching like polarities between the electromagnets and the neodymium magnet, haptic force feedback was achieved.

FingerFlux [4] is an output technique which generates near-surface haptic feedback in interactive tabletops. It combines electromagnetic actuation with a permanent magnet attached to the user's hand. FingerFlux provides enhanced features like, feel the interface before touching, attraction and repulsion, development of applications such as reducing drifting, adding physical constraints to virtual controls, and guiding the user without visual output.

Texture Display Mouse[13] is a haptic offers the capability of displaying properties such as patterns, gratings, and roughness. The array can represent micro-scale shapes

with various surfaces, such as gratings, grooves, patterns, shapes of icons, and Braille, and provides the user with cutaneous stimuli. Tactile Explorer [5] is a device which provides access to computer information for the visually handicapped people using tactile sensations. The tactile mouse resembles a regular computer mouse, but differs in having two tactile pads on top that have pins that move up and down.

Microsoft explorer mouse [6] is a commercially available mouse implementation which combines a light haptic sensation. Haptic-feedback, in the form of vibration through the touch-sensitive strip, indicates which one of the three scrolling speeds has been selected. Both Tactile Explorer and Microsoft tactile mouse are mouse implementations combined with Haptic. It supports enhanced haptic interactions. However, operations and sensations are limited to the device surface. Furthermore, the haptic actuation is limited to a small area of the device surface.

3 System Overview

Magnetic field based Near Surface Haptic and Pointing Interface contains three modules. They are Sensing System, Software Interface Driver and Actuation System. These three parts are described in the following sections.

3.1 Sensing System

Sensing mechanism basically concentrates the tracking of the neodymium magnet using an array of Hall Effect sensors. Hall Effect sensors array is a 2D implementation, as can be seen in the figure 1. The distance between two Hall Effect sensors of the array is 2cm along the X and Y directions. Once the neodymium magnet is moving on top of the sensors grid, sensors output DC voltage values which can be converted in to a digital using the ADC converters. We have implemented a scale to measure the strength of the magnetic field detected by the sensors. The scale contains values from 0 to 1024, where first half of the range (0-512) is used when a sensor sensed a magnetic field produced by the north pole of the neodymium magnet and second half of the range is used (513-1024) when a sensor sensed a magnetic field produced by the south pole of the neodymium magnet.

We have recorded the outputs of a single Hall Effect sensor while changing the position of the neodymium magnet along X,Y, and Z axis. There, we have discovered that the sensor reading is at the highest reading for the Z axis in between the range from 0mm to 20mm. Therefore we had to lift the device surface 2cm above the sensors grid to obtain dynamically changing readings with the distance. We were able to track 3D position of the neodymium magnet which attached to the fingertip, up to 3 cm above the device surface. The mathematical model used to track the position is mentioned in the section 3.2.

3.2 3D Localization Algorithm

To move the mouse pointer in the screen we have used the Neodymium magnet attached in the finger tip. By moving the particular finger user was able to change the position of the mouse pointer. It is crucial to detect the correct position of The neodymium magnet over the time because failing to localize the magnet interrupt the

continuous movement of the mouse pointer. In our previous work [14] we have presented a 2D localization algorithm to determine the position of the magnet. However, its accuracy was limited to surface of the device. Once the neodymium magnet is some millimeters above the surface at some points the system has failed to detect the magnet. To overcome the shortcomings of the previous algorithm and looking for the possibilities to apply 3D gestures we have developed a new 3D localization algorithm.

We have done a preliminary experiment to investigate the variation in the magnetic field strength vs. the distance of all three axes and determine the strength of the magnetic field. This experiment was conducted by positioning the neodymium magnet on top of the Hall Effect sensor and measuring output readings at various distances in all three axes and results are shown in Table 1.

Table 1. Hall effect sensor readings for the X, Y and Z axis

Distance	X	Y	Z
0	963	960	955
2.5	933	933	931
5	895	893	898
7.5	840	839	835
10	765	760	757
12.5	672	675	665
15	635	635	636
17.5	597	588	590
20	563	565	564
22.5	543	542	543
25	528	527	523
27.5	519	518	520
30	513	514	516

Table 2. Expressions to determine the distance between the hall effect sensor and neodymium magnet based on the hall effect readings

Distance(mm)	M	C	Expression
0-2.5	-12	963	$Y = -12X + 963$
2.5-5	-15.2	971	$Y = -15.2X + 971$
5-7.5	-22	1005	$Y = -22X + 1005$
7.5-10	-30	1065	$Y = -30X + 1065$
10-12.5	-37.2	1137	$Y = -37.2X + 1137$
12.5-15	-14.8	857	$Y = -14.8X + 857$
15-17.5	-15.2	863	$Y = -15.2X + 863$
17.5-20	-13.6	835	$Y = -13.6X + 835$
20-22.5	-8	723	$Y = -8X + 723$
22.5-25	-6	678	$Y = -6X + 678$
25-27.5	-3.6	618	$Y = -3.6X + 618$
27.5-30	-2.4	585	$Y = -2.4X + 585$

According to the results shown in table 1, it is clear that sensor reading values are following nonlinear curves but along the X,Y and Z axis the readings are approximately the same.

As shown in figure 2, the distances to the neodymium magnet from a sensor can be illustrated as circles or a spheres [12]. Data from a single sensor helps to narrow the possibility of the neodymium magnet’s position down to a large area of sphere around the particular sensor. Adding data from a second sensor narrows position down to the region where two spheres overlap. Adding data from a third sensor provides two possible points where magnet can be exists. However, in this setup we placed all the sensors such as Z=0 and coordinates of the two possible points becomes (x,y,z) and (x,y,-z). Since the magnet is placed on top of the surface we have the freedom to select (x,y,z) as the correct position.

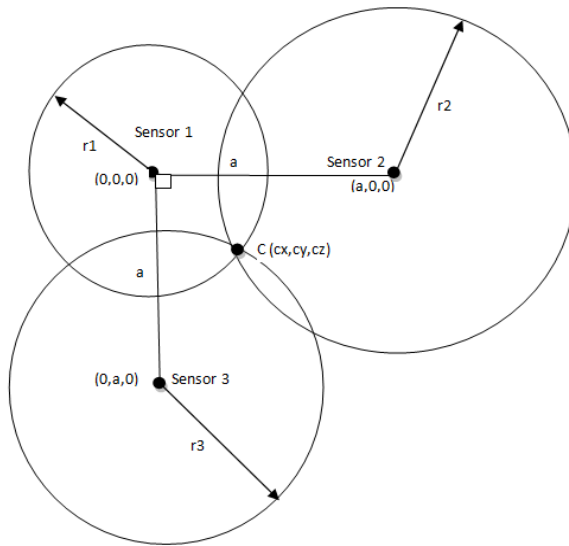


Fig. 2. Distance from the sensors to the neodymium magnet can be calculated using the output voltage of the sensors

Further, to simplify the calculations, the equations are formulated as the location of the sensor which forms a right angle triangle (Sensor 1) is at the origin, and one other is on the x-axis (Sensor 2). By using the general equation for spheres,

$$r^2 = x^2 + y^2 + z^2 \tag{1}$$

We could write the expressions for the S1, S2, and S3 as follows

$$r_1^2 = x^2 + y^2 + z^2 \tag{2}$$

$$r_2^2 = (x - a)^2 + y^2 + z^2 \tag{3}$$

$$r_3^2 = x^2 + (y - a)^2 + z^2 \quad (4)$$

By subtracting the third equation from the second equation we can obtain a solution for x,

$$x = \frac{r_1^2 - r_2^2 + a^2}{2a} \quad (5)$$

We assume that S1 and S2 spheres intersect in more than one point. In this case substituting the equation for x back into the equation for the S1 produces the equation for a circle, the solution to the intersection of the first two spheres,

$$\begin{aligned} y &= \frac{r_1^2 - r_3^2 - x^2 + (x - a)^2 + a^2}{2a} \\ &= \frac{r_1^2 - r_3^2 - 2ax + 2a^2}{2a} \\ &= \frac{r_1^2 - r_3^2}{2a} + (a - x) \end{aligned} \quad (6)$$

By rearrange the formula for the first sphere to find the z-coordinate,

$$z = \pm \sqrt{r_1^2 - x^2 - y^2} \quad (7)$$

After finding the solution relative to the point which causes a right angle triangle (sensor 1), we have transformed the position of the neodymium magnet to the original three dimensional Cartesian coordinate system using the coordinates of S1.

3.3 Software Interface Driver

This driver accepts the row sensor values converted to digital from the microcontroller (Arduino Mega 2560) of the Hall Effects sensors grid as the input. These sensor values are sorted in the descending order and if the magnet is North Pole downwards, software searches for the positions of the sensors in the grid where it received the maximum readings. Sensors which are nearest to the neodymium magnet, output the maximum values. Based on those intensity values relative distance to the neodymium magnet from the nearest three sensors are calculated using the localization algorithm in the Section 3.3. By finding the position of the neodymium magnet and comparing it with the next position, relative X,Y displacement can be calculated. Then these relative displacements are mapped to the last coordinates of the mouse cursor position and moves the cursor to a new X,Y location.

In the case of identified mouse commands, firstly, driver identifies the neodymium magnet which is placed South Pole downwards by reading the digitally converted values. If the magnet is South Pole downwards, software driver searches for the three minimum sensor reading values and determines the coordinates of those sensors. Then, the distance to the neodymium magnet from each sensor is calculated and its

position is determined using the 2D Trilateration based technique presented in our earlier paper [14]. The movement path of the neodymium magnet is tracked and if the path follows the gestures defined for the mouse commands, the driver activates the appropriate commands. As the final step, it updates Electromagnet controller circuit about the necessary vibration pattern which would eventually provide the user with the vibration feeling.

3.4 Actuation System

Haptic Mouse provides attraction and repulsion sensations by changing the polarity of the electromagnets. Polarity is changed by swapping the positive and negative voltage supply to electromagnets using a controller circuit. When the neodymium magnet worn on the finger tips and the electromagnet array positioned in the opposite polarity (N – S or S - N) users feel an attraction towards the device surface. Users feel the repulsion sensation when those magnets are in like polarity (S - S or N-N) positions.

Vibration sensations are provided by setting up neodymium magnet and magnetic array in a like polarity position and then rapidly switching on and off the electromagnetic array in certain frequencies. This rapid switching on and off dynamically changes the magnetic field it produces and affects the static magnetic flux developed by the neodymium magnet worn on the finger tips. While electromagnet is switched off neodymium magnet comes down but when the electromagnet is switched on it rises and this is felt by the user as a vibration. In the case of sensing the shapes, driver software keeps a selected vibration pattern until the user move the mouse cursor on top of the interested object in the screen. Once the cursor is moved away from the object boundary, driver sends commands to the microcontroller of the electromagnet controller circuit to change the output frequency.

This part of the system is made with six electromagnets, Magnet controller circuit and Arduino based microcontroller. As the total power required by the electromagnets array is high at 6V and 13A [3], it becomes necessary to control the power supplied to the electromagnets via a relay circuit. To address this, the relay circuit acts as a mechanism that is able to switch on a much larger power to drive the electromagnets. For this power up electromagnets, six N-Type MOSFET [8] were used, one for each electromagnet.

4 Results

We have evaluated the accuracy of the 3D localization algorithm discussed in the section 3.2. Hall Effect sensor grid used in this device is a 4*3 array (4 sensors along the X axis and 3 sensors along the Y axis). The space between two Hall Effect sensors was allocated as 100 pixels. Therefore, all the sensor values recorded are represented as X,Y coordinates (0-300 in X axis and 0 to 200 in Y axis). This experiment was conducted by moving the neodymium magnet on top of the device surface along four straight lines which are randomly picked. $Y=80$, $Y=(3/5) X$, $Y= -(2/3)X +200$, $X=170$ are those lines. We have used two rulers and a digital Vernier caliper to place the Neodymium magnet in the correct position. The result of the experiment is illustrated in the “Fig.3”.

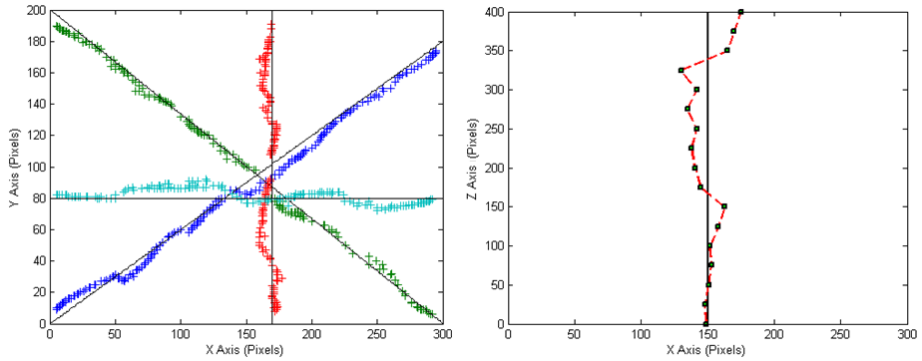


Fig. 3. Accuracy test of the X, Y and Z axes

According to Figure 3, the sensors were capable to detect the motion of the neodymium magnet in near linear fashion on the surface. Further, sensors managed to detect the position more than 90% of points with less than 5% of error. This line does not reflect the movement of the mouse cursor. Mouse cursor position is calculated by adding the difference of the X,Y displacement between two neodymium magnet position readings. Therefore, the accuracy of the movement of the mouse cursor was further improved by cancel out differences which passes certain threshold value. Figure 3 also shows the position detection readings of the neodymium magnet along the Z axis from the device surface to the 4cm above. Sensors were able to track the position near accurately; however, increasing the height from the surface level along the Z axis sensing module loses the accuracy. This may be due to two reasons, the limitations of the Hall Effect sensors and inaccuracies of the 3D localization algorithm.

5 Conclusion and Future Work

To conclude, in this paper we have presented a new type of computer interface which provides basic pointing interface functionalities with near surface haptic feedback. Haptic feedback can be felt up to 6 cm of height and pointing functionality is worked up to 3cm above the device surface. Implementing variable friction for haptic interface using this technology will be an interesting research topic in future since variable friction has not been implemented for touch sensitive haptic feedback systems. TeslaTouch [11] is the closests excusion of such haptic display . This device can be improved as an interface for visually handicapped who rely mostly on touch sensation. In orderto improve to this level of proficiency, this system is required to minimize the size of the electromagnets and increase the density of electromagnets packed in the electromagnets array which will provide a better resolution. This device could also be improved as an easy learning tool for children, which can be used to draw some basic shapes or characters that will enhance the interactive enjoyment.

Acknowledgment. This research is carried out under CUTE Project No. WBS R-7050000-100-279 partially funded by a grant from the National Research Foundation (NRF) administered by the Media Development Authority (MDA) of Singapore.

References

1. Virvou, M., Kabassi, K.: Reasoning about users' actions in a graphical user interface. *Hum.-Comput. Interact.* 17(4), 369–398 (2002)
2. Karunanayaka, K., Koh, J.T.K.V., Naik, E.B., Cheok, A.D.: Hall effect sensing input and like polarity haptic feedback in the liquid interface system. In: Keyson, D.V., Maher, M.L., Streitz, N., Cheok, A., Augusto, J.C., Wichert, R., Englebienne, G., Aghajan, H., Kröse, B.J.A. (eds.) *AMI 2011. LNCS*, vol. 7040, pp. 141–145. Springer, Heidelberg (2011)
3. Arduino, <http://www.arduino.cc/en/>
4. Weiss, M., Wacharamanotham, C., Voelker, S., Borchers, J.: FingerFlux: near-surface haptic feedback on tabletops. In: *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, UIST 2011, Santa Barbara, California, USA, October 16-19*, pp. 615–620. ACM, New York (2011)
5. Tactile Explorer, <http://www.tactile-world.com/>
6. El Saddik, A., Orozco, M., Eid, M., Cha, J.: *Haptics Technologies Bringing Touch to Multimedia*. Springer Series on Touch and Haptic Systems (2011)
7. Magnetech Corporation, http://www.magnetechcorp.com/Opposite_Pole.htm
8. MOSFETs, <http://www.falstad.com/circuit/e-nmosfet.html>
9. Engelbart, D.C., et al.: *A Research Center for Augmenting Human Intellect (demonstration)* Stanford Research Institute, Menlo Park (1968), <http://sloan.stanford.edu/MouseSite/1968Demo.html>
10. El Saddik, A.: The potential of haptics technologies. *IEEE Instrumentation & Measurement Magn.* 10, 10–17 (2007)
11. Bau, O., Poupyrev, I., Israr, A., Harrison, C.: TeslaTouch: Electro-vibration for Touch Surfaces. In: *Proceedings of UIST 2010*, pp. 283–292. ACM (2010)
12. Manolakis, D.E.: Efficient Solution and Performance Analysis of 3-D Position Estimation by Trilateration. *IEEE Trans. on Aerospace and Electronic Systems* 32(4), 1239–1248 (1996)
13. Kyung, K.-U., Kim, S.-C., Kwon, D.-S., Srinivasan, M.A.: Texture display mouse kat: Vibrotactile pattern and roughness display. In: *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 478–483 (2006)
14. Karunanayaka, K., Siriwardana, S., Edirisinghe, C., Nakatsu, R., Gopalkrishnakone, P.: Haptic Mouse - Enabling Near Surface Haptics in Pointing Interfaces. In: *The Sixth International Conference on Advances in Computer-Human Interactions, ACHI 2013, Nice, France*, pp. 336–341 (March 2013)

Use of Reference Frame in Haptic Virtual Environments: Implications for Users with Visual Impairments

Ja Young Lee, Sangwoo Bahn, and Chang S. Nam

North Carolina State University, Raleigh, NC, USA
{jlee47, csnam}@ncsu.edu, panlot@gmail.com

Abstract. Reference frame is key in explaining the relationship between two objects. This paper focused on the orientation parameter of a reference frame in use of projective spatial terms, and its use by visually impaired participants using a haptic device to explore a haptic virtual environment. A total of nine visually impaired participants between 12 and 17 years of age participated in this study. After exploring the 3D virtual environment with a haptic device, participants answered questions about the frame they had utilized. Overall results indicated that the participants used relative frame of reference slightly more than the intrinsic frame of reference. This inclination was especially clear when both the target object and the reference object were on the horizontal plane. Only when objects were on horizontal plane but intrinsically vertical to the reference object, the intrinsic frame of reference was preferred. We also found evidence that participants used a reflective subtype of the relative frame, and vertically aligned objects were easy to be perceived with the relative reference frame. We concluded that the virtual environment and haptic input had influence on the result by separating the user from the computer, only allowing one point of contact. Thus it would be possible to apply the result of this study to the development and assessment of assistive technology for people with visual impairment, especially in regard to how spatial information between the systems and the user is communicated.

Keywords: Reference frame, relative frame, intrinsic frame, projective spatial terms, visual impairments.

1 Introduction

When people explain a spatial relationship between two objects, especially the direction from one to the other, they use projective spatial terms such as ‘left’, ‘front’, ‘above’, and so on. Use of such terms is based on different frames of reference. Imagine you are facing a computer screen, with a coffee cup placed on the right hand side of the screen. The coffee cup is obviously ‘to the right of the screen’, but at the same time, it is ‘to the left of the screen’ from the perspective of the screen. The difference between these two explanations comes from different frames of reference that each sentence buys. The first sentence implies ‘relative’ frame of reference, and the second sentence implies ‘intrinsic’ frame of reference.

There is a wealth of distinctions across many disciplines that explicitly use the term ‘spatial frames of reference’. In this paper, we used the three linguistic frames of reference defined by Levinson [1].

- Intrinsic frame of reference: The coordinate system that uses features of the reference object to explain the location of the target object.
- Relative frame of reference: The coordinate system that is established by the position and functional-spatial structure of an additional entity, usually an egocentric viewpoint.
- Absolute frame of reference: The coordinate system that uses features of the environment, such as gravity, cardinal directions, or landmarks.

Appropriate implementation of a reference frame is associated with the viewpoint people take when they create a spatial mental map. It is especially important for visually impaired people because spatial language acts as an alternative to visual information. Many papers indicate that vision is not the only modality utilized in creating mental maps; the nature of spatial images is supramodal [2] [3] [4]. In many studies, people were able to convert verbal descriptions into mental representations that are similar to and function equivalently with the mental representation derived from visual experience [5] [6] [7].

One point to consider is that the cognitive processes of people with visual impairment may differ from that of sighted people. Levinson [1] suggested that there might be differences between visually impaired people and normally sighted people in spatial language, largely due to the dependency of language-space interaction on former experience, rather than given priory. A review by Cattaneo et al. [8] reported that cognitive mechanisms, or mental processes, are strongly affected by the nature of perceptual input on which people commonly rely on.

In particular, there is evidence that people with visual impairment prefer the intrinsic frame of reference. Struiksmma and colleagues [9] conducted an experiment using projective spatial terms in order to observe the preference of reference frame among blind people, low-vision people, and sighted people groups. The results indicated that the blind group showed a clear preference for the intrinsic frame, when judging spatial relation in the horizontal plane. According to the study of Postma et al. [10], sighted people tended to prefer the absolute reference frame in order to point out the locations of objects, while blind people would rather use the intrinsic reference frame.

Nevertheless, there is no previous research on reference frames in haptically enhanced virtual environments, despite the fact that haptic virtual environments can enhance learning of people with visual impairment as an assistive technology. Lahav and Mioduser [11] tested a virtual environment with a haptic device in order to provide visually impaired people with prior spatial information on unexplored space. In such a situation, we assume that feedback with accurate spatial language will enhance the usability and reliability of a system.

This paper focused on exploration of a haptically enhanced virtual environment and investigated which reference frame people with visual impairment prefer in order to perceive the spatial relationship between two objects. The results of the present study may serve as a basis for the study of spatial language in haptic virtual

environments. In addition, it will suggest ways to minimize system-to-user or user-to-user communication. In the following sections, we will introduce our system and experiment procedure, and discuss the results we obtained.

2 Methods

2.1 Participants

We recruited participants from a local school for the blind. Nine participants were included with varying degrees of blindness; three participants were totally blind, one was nearly totally blind, and five were partially blind. The ages of participants were between 12 and 17, and in middle or high school (between 6th and 11th grade). Three participants were male, and six participants were female.

2.2 System

New software was developed based on Novint Falcon SDK to create the 3D experimental virtual environment. The system enabled arrangement of objects of desired shapes and sizes in a 3D space. When stimuli were arranged in the 3D space, users could detect them with a Novint Falcon haptically enhanced 3D touch controller.

2.3 Stimuli

Two stimuli were utilized: a target object and a reference object. The target object was a ball (represented as a sphere), and the reference object was a car (represented as a cube). These objects were placed in the virtual 3D space, where no wall, floor, or ceiling is detected. To create an orientation cue for the reference object, we provided a miniature of the car (10cm×5cm×3cm) fixed on a wooden plate in the real world, whose configuration was altered based on a layout of each trial. Experimenters verified that participants understood the directionality of the toy car (i.e., which side is front, left, and above) before the experiment.

During the experiment, the participants detected overall layout of stimuli in the virtual world by controlling the Falcon device with their dominant hand. At the same time, they could feel the shape and direction of the reference object (miniature car).

2.4 Procedure

Throughout the experiment, participants were seated on a chair in front of a computer desk. The Falcon device was located on the desk, on the side of each participant's dominant hand. The wooden plate holding the miniature of the reference object was placed in front of the participants. The computer screen was turned to the experimenter and away from the subject to prevent partially blind participants from seeing the screen.

We used the terms ‘front’, ‘left’, and ‘above’ to represent three axes of the 3D space, and two different frames of reference: intrinsic and relative/absolute. The absolute frame of reference goes together with the relative frame of reference in this case because gravity determines the vertical axis of body posture, a basis of the relative frame of reference.

Each cell in the Table 1 below indicates each trial. For example, the trial in the second column of the first row, layout 2, represents the relationship between objects that can be either “the ball is on the left of the car” (based on relative/absolute frame) or “the ball is in front of the car” (based on intrinsic frame). We removed three cells that use same terms for both reference frames (grey cells) because of redundancy.

Table 1. Layout

		Front	<i>Relative</i> Left	Above
<i>Intrinsic</i>	Front	Layout 1	Layout 2	Layout 3
	Left	Layout 4	Layout 5	Layout 6
	Above	Layout 7	Layout 8	Layout 9

For each layout, the participants were given enough time to explore the layout of the virtual environment with the haptic device. Then, they were asked to judge the truth or falsehood of two statements describing the relationship between objects. For instance, in the example mentioned above, participants were to give answers to both “the ball is on the left of the car” (true based on relative/absolute frame) and “the ball is in front of the car” (true based on intrinsic frame).

3 Result and Discussion

Each participant judged truth and falsehood of two statements describing each of six layouts. Hence, we could have 108 boolean data points in total. If one used relative reference frame, he or she would have answered true to the relative frame based projective spatial term (e.g., in layout 2, a participant would say ‘true’ to the term ‘left’ if relative frame is used, but say ‘false’ to the term ‘in front’). If they used intrinsic reference frame, they would have answered true to the intrinsic frame based term (e.g., in layout 2, a participant would say ‘true’ to the term ‘in front’ if intrinsic frame is used, but say ‘false’ to the term ‘left’)

On average, 55.56% among all responses were answered true to the relative frame, whereas 46.30% answered true to the intrinsic frame. The sum of these two is not equal to 100% because only 64.81% used one reference frame at a time. Among those who used single frame at a time, 57.14% used the relative frame and 42.86% used the intrinsic frame. Figure 1 shows this tendency with the proportion of participants saying ‘true’ to both frames (18.52%), and ‘false’ to both frames (16.67%).

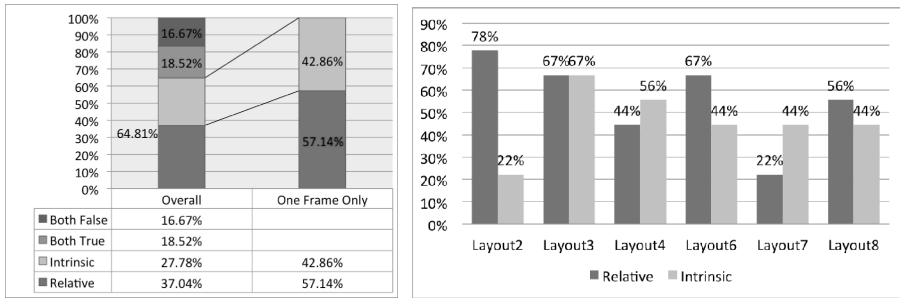


Fig. 1. Overall results (left) and result from each layout (right)

The graph on the right side of the Figure 1 shows the results from each layout. Among six layouts, the participants tended to use relative reference frame in the case of layout 2, 6, and 8, while they tended to use intrinsic frame in layout 4 and 7.

To analyze the results for each frame, we first combined the results focusing on relative reference frame. For instance, if we combine layouts 4 and 7, it is the case when ‘front’ indicates relative reference frame, regardless of the intrinsic reference frame. In the same sense, we integrated layouts 2 and 8, and layouts 3 and 6. As a result, it turned out that participants tended to use intrinsic frame in the first combination where ‘front’ indicates relative reference frame. Other trials where ‘left’ and ‘above’ indicates relative reference frame, however, the participants tended to use the relative frame rather than the intrinsic frame.

If we instead disregard relative frame, layouts 2 and 3 can tied together where ‘front’ represents intrinsic frame. Similarly, layouts 4 and 6, and layouts 7 and 8 fell into the same category. In this case, participants tended to use the intrinsic frame when ‘above’ represents intrinsic frame (the combination of layout 7 and 8). Yet, we could not find clear preference to the intrinsic frame, as the difference was too small.

Figure 2 demonstrates the results described above. This analysis also suggests that the relative frame was dominant in general. The graph with error bars shows that there was no case when the intrinsic frame outperformed the relative frame. (The ambiguous ‘front’ case in the left graph is discussed in Figure 3.)

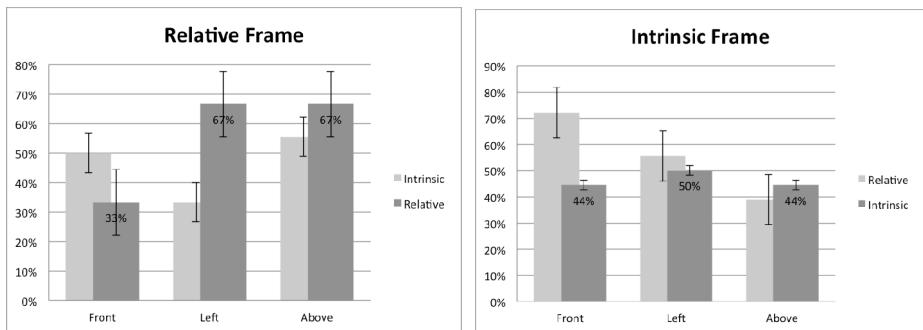


Fig. 2. Percentage of selecting each reference frame for a specific term

We also investigated the results in terms of spatial planes: horizontal and vertical. We combined layout 2 and layout 4 together, as both layouts are explainable with horizontal terms, either ‘front’ or ‘left’, regardless of the frame (HH). Likewise, we integrated layout 3 and layout 6. In both cases, relative frame corresponds to the vertical term ‘above’, and intrinsic frame corresponds to the horizontal term ‘left’ or ‘front’ (VH). Layout 7 and layout 8 used the term ‘left’ or ‘front’ for relative frame and ‘above’ for intrinsic frame (HV). The graph on the left side of Figure 3 outlines the result of this integration. The participants showed preference to the relative frame when the two objects were on the horizontal plane, which is a canonical situation. The tendency was similar when the target object was relatively above, even though it is a non-canonical situation. However, this trend was marginally inverted when the objects were on the horizontal plane but the target object was intrinsically above the reference object. It suggests that the proportion of people who think in the perspective of the reference object increases when it is a non-canonical situation and the objects are not aligned on the vertical plane.

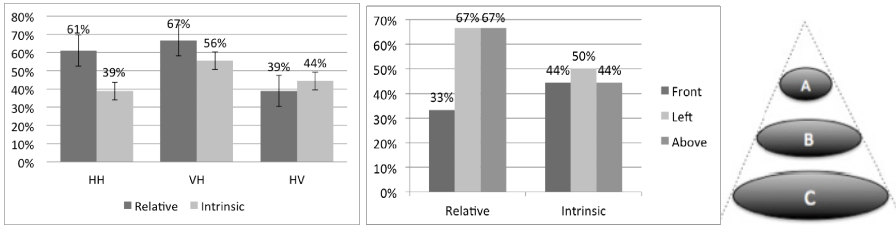


Fig. 3. Percentage for choosing either reference frame, under three types of layout: two objects are on the horizontal plane and intrinsically horizontal (HH), two objects are on the vertical planes but intrinsically horizontal (VH), and two objects are on the horizontal plane but intrinsically vertical (HV) (left); Percentage of ‘true’ answers for each reference frame under a specific term (middle); the ‘front’ concept (right)

The case where participants answered ‘both false’ or ‘both true’ led to an interesting result. Out of five participants who answered false to both frames, four made ‘both false’ answer to the layout 7, where the target object was relatively front but intrinsically above. One possible explanation is that the participants thought the target object was relatively ‘behind’ the reference object. The graph in the middle of Figure 3 shows the percentage of ‘true’ to each frame, which is part of Figure 2. We can find a noticeable outlier on the far left side; the ratio of ‘true’ answer to the ‘front’ was significantly lower than other cases of the relative frame. One possible explanation is that the participants did not perceive our intended ‘front’ as ‘front’. In the experiment, we assumed that the target object is in front of the reference object when it is further away. The diagram on the right side of Figure 3 explains this; what we intended was that ‘A is in front of B’. However, the result suggests that participants might have perceived ‘C is in the front of B’.

Such difference yields subtypes of the relative reference frame: translation and reflection. When a target object beyond a reference object is considered to be ‘in front of’ the reference object, it is of the translation subtype; when a target object between a

perceiver is considered to be ‘in front of’ the reference object, it is of the reflection subtype. Generally, different language affects customs of such subtypes, and English entails the reflection subtype (Cox 1981, Levinson 2003, Bender et al. 2005). The inclination of English speakers to rotate their body orientation by 180 degrees when they use the relative frame of reference may have caused the interesting outcome.

Furthermore, out of ten cases where the participants answered ‘true’, six cases were from layout 3 and 6, or VH case (vertical term stands for relative frame, and horizontal terms stand for intrinsic frame). It implies that absolute above is easy to perceive and stable regardless of the intrinsic orientation of a reference object. It partially coincides with previous research from Struiksma et al. [9], where the blind people showed relatively low bias to the intrinsic frame of reference when the objects were aligned vertically.

Our study results showed more familiarity with the relative frame of reference in general, which does not match with the previous experiments where relatively large number of intrinsic reference frame responses took place. First, it may be due to the virtual environment providing fewer sensory cues than most physical environments, causing lower presence in the virtual environment [12]. Since the virtual environment lacks sensory channels, it could be hard for the users to think in the perspective of the reference object, or intrinsic frame. If this is a correct explanation, we can also suggest the reference frame test as a tool to assess the level of engagement in the virtual reality, breaking boundaries of traditional subjective questionnaire methods [13].

Second, the ambiguity caused by one-point movement might have influenced the perception. The haptic inspection of configurations required one point exploration in this study; with Novint Falcon, the participant could only touch the virtual objects with a one-point cursor, on the contrary to the ordinary haptic situations where they normally employ two hands and ten fingers. Subjective inspection also showed that the surface area of objects they actually touched was relatively small. It is possible that the limited touch caused confusion and resulted in participants using the relative frame of reference, which is fairly easy to apply in that it does not require any mental rotation.

4 Conclusion

The participants with visual impairments used the relative reference frame in preference to the intrinsic reference frame when they perceived the objects in a 3D virtual environment. This tendency was constant, except for the case where the target object is intrinsically above in relation to the reference object. The results dissent from the studies prior to this research, where people with visual impairment mostly preferred the intrinsic frame of reference. As the former studies were conducted in the real world using real objects, the virtual environment might have produced different aspects of perception, in terms of framing spatial relationships. The characteristic of haptic exploration could be one other factor that caused people to use the relative frame as well.

To improve the precision of results, a larger sample size with more encompassing statistical analysis is required. Furthermore, we cannot ignore cultural effects.

Considering the participant groups with different cultural backgrounds in different papers, we will be able to make stronger suggestions if we compare the results to those of a group of sighted people within a same culture and age group.

References

1. Levinson, S.C.: Frames of reference and Molyneux's questions: cross-linguistic evidence. In: Bloom, P., Peterson, M.A., Nadel, L., et al. (eds.) *Language and Space*, pp. 109–169. MIT Press, Cambridge (1996)
2. Carpenter, P., Eisenberg, P.: Mental Rotation and the Frame of Reference in Blind and Sighted Individuals. *Attention, Perception, & Psychophysics* 23, 117–124 (1978)
3. Jones, B.: Spatial Perception in the Blind. *Br. J. Psychol.* 66, 461 (1975)
4. Struiksma, M.E., Noordzij, M.L., Postma, A.: What is the Link between Language and Spatial Images? Behavioral and Neural Findings in Blind and Sighted Individuals. *Acta Psychol.* 132, 145–156 (2009)
5. Avraamides, M.N., Loomis, J.M., Klatzky, R.L., et al.: Functional Equivalence of Spatial Representations Derived from Vision and Language: Evidence from Allocentric Judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 30, 801–814 (2004)
6. Denis, M., Zimmere, M.: Analog Properties of Cognitive Maps Constructed from Verbal Descriptions. *Psychological Research* 54, 286–298 (1992)
7. Loomis, J.M., Lippa, Y., Klatzky, R.L., et al.: Spatial Updating of Locations Specified by 3-D Sound and Spatial Language. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 28, 335–345 (2002)
8. Cattaneo, Z., Vecchi, T.: Supramodality Effects in Visual and Haptic Spatial Processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 34, 631–642 (2008)
9. Struiksma, M.E., Noordzij, M.L., Postma, A.: Reference Frame Preferences in Haptics Differ for the Blind and Sighted in the Horizontal but Not in the Vertical Plane. *Perception* 40, 725–738 (2011)
10. Postma, A., Zuidhoek, S., Noordzij, M.L., et al.: Differences between Early-Blind, Late-Blind, and Blindfolded-Sighted People in Haptic Spatial-Configuration Learning and Resulting Memory Traces. *Perception* 36, 1253–1265 (2007)
11. Lahav, O., Mioduser, D.: Exploration of Unknown Spaces by People Who are Blind using a Multi-Sensory Virtual Environment. *Journal of Special Education Technology* 19, 15–24 (2004)
12. Biocca, F., Kim, J., Choi, Y.: Visual Touch in Virtual Environments: An Exploratory Study of Presence, Multimodal Interfaces, and Cross-Modal Sensory Illusions. *Presence: Teleoperators & Virtual Environments* 10, 247–265 (2001)
13. Slater, M.: How Colorful was Your Day? Why Questionnaires Cannot Assess Presence in Virtual Environments. *Presence: Teleoperators & Virtual Environments* 13, 484–493 (2004)
14. Cox, M.: Interpretation of the Spatial Prepositions' in Front of' and 'Behind'. *International Journal of Behavioral Development* 4, 359–368 (1981)
15. Levinson, S.C.: *Space in language and cognition: Explorations in cognitive diversity*, vol. 5. Cambridge University Press (2003)
16. Bender, A., Bennardo, G., Beller, S.: Spatial Frames of Reference for Temporal Relations: A Conceptual Analysis in English, German, Tongan, pp. 220–225 (2005)

Behavioral Characteristics of Users with Visual Impairment in Haptically Enhanced Virtual Environments

Shijing Liu, Sangwoo Bahn, Heesun Choi, and Chang S. Nam

North Carolina State University, Raleigh, NC, USA
{sliu14, sbahn, heesun_choi, csnam}@ncsu.edu

Abstract. This study investigated behavioral characteristics of users with visual impairments and tested effect of factors regarding the layout of virtual environments (VEs). Various three-dimensional (3D) VEs were simulated with two different factors: number of objects and layout type (random, symmetric). Using a Novint Falcon haptic device, users with visual impairments were required to complete an object recognition task in 3D VEs with different levels of number of object and layout. The characteristics of their movements (speed, applied force, location, direction, etc.) were recorded, and participants evaluated perceived difficulty after they completed each trial. We analyzed their recorded movements and their rating on perceived difficulty. Results showed that 1) number of objects in 3D VE had significant impact on visually impaired users' behavior; 2) different layout had not showed significant influence on their movement; 3) increased number of objects in 3D VE made the task more difficult; 4) visualized results implied that different users had significant different behavior preference in the same 3D VE. It is expected that the results of this study can improve behavioral understanding of users with visual impairments and guidance for assistive technology development for users with visual impairments.

Keywords: Haptic, 3D virtual environment, behavioral pattern.

1 Introduction

Spatial cognition is essential to represent, organize, understand, and navigate the environment, to attend to specific objects, to manipulate objects, and to communicate information about objects and the environment to others (Spence et al., 2010). Spatial reasoning is often considered to be one of the fundamental abilities for survival as well as everyday activities such as driving, way-finding, and learning. There have been extensive studies investigating human behavior and spatial performance in various spatial tasks both in real and virtual environments (VE).

Many past studies focused on the visual channel as an interaction modality to investigate spatial performance in VE and information processing in navigation. The visual sensory system has been used as a primary channel for humans in most virtual spatial activities because visual modality can deliver the most accurate and rich

spatial information. However, other types of modalities can be beneficial as redundant or supplemental channels to receive spatial information and control spatial movement in VE.

Previous research has suggested that the tactile modality can be a strong alternative sensory channel for visually impaired users in both real and virtual environments. Previous studies have found that a multisensory VE (auditory and tactile) supported development of spatial representation of virtual map and improved navigation performance in real environments of blind people (Lahav et al., 2003), and interaction tools with tactile and auditory modalities help people with visual impairment with tasks in web or graphic programs (Roth et al., 2000). It also has been found that a tactile map has superior effectiveness over a verbal description of the area on development of spatial knowledge in an unfamiliar environment (Espinosa et al., 1998). Past research has also examined various types of tasks and systems using tactile interface for people with visual impairment including haptic data visualization system design (Fritz et al., 1999), computer-based haptic graphs design (Yu et al., 2000), and haptic enhanced user interface in science learning (Li et al., 2011).

It has been suggested that an investigation of human behavioral characteristics can provide a fundamental understanding about how humans interact with their environment as well as impacts of significant environmental factors on performance. One past study suggested that movement is associated with the way in which people experience the VE (Särkelä et al., 2009). However, there has not been enough attention to investigate in-depth behavioral characteristics of people with visual impairment while they navigate and receive feedback in a haptically enhanced VE. Another major limitation in previous literature is that there have been only limited studies examining the influence of specific environmental or physical features of a haptic interface. Little is known about the effect of specific environmental factors in VE on behavior patterns in spatial cognitive and navigation performance in people with visual impairment.

The current study aimed to investigate the impact of different environmental characteristics in a 3D VE in which visually impaired participants control navigation and receive feedback through their tactile sensory system. Numbers of objects and types of object layout (random or symmetric) were controlled in virtually generated spatial environments to examine how these manipulated environmental characteristics influence behavioral patterns and perceived difficulty in an object recognition task of people with visual impairment.

2 Methods

2.1 Variables

We employed two elements that can characterize the environment of a virtual space: number of object (three levels, $N=4, 8, 12$) and layout type (random, symmetric). In order to minimize effects of other influence factors such as object size and space density, we randomized these other factors. We used task completion time and completion ratio for measuring performance, and we also measured the cursor's

location, speed, applied forces and direction (30 times per one second) to quantify users' actual behavior.

2.2 Apparatus

For the experiment, six haptically enhanced 3D VEs were developed for use with a Novint Falcon. The Falcon is a three dimensional haptic device, which can control the on screen cursor, enabling computer interaction. Additionally, it is able to receive and provide feedback between users and computer. For each experiment station, we set up a Falcon haptic device and one computer with two monitors, one for experimenter control and the other one for observation of user movement. Fig. 1 shows the haptic device and experiment environment.

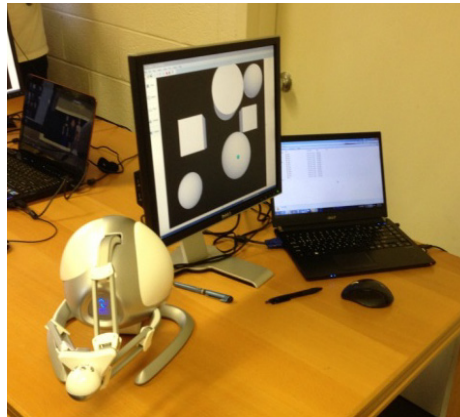


Fig. 1. The device used in the experiment

2.3 Participants

Six users with visual impairment were recruited from a school for blind students (age: 15 – 20 ($M = 17.5$, $SD = 1.87$), gender: one male, five females). Five were partially blind, and one was almost totally blind. They spent an average of 2.88 hours per week on a computer. Four had prior experience with haptics of some kind (haptic devices, force/vibration feedback from electronic devices, etc.).

2.4 Procedures

All participants were required to complete a demographic survey before the beginning of the experiment. Then, a training session was provided to them. During the training session, experimenters guided exploration of the 3D VE to ensure recognition of two types of objects (cube and sphere). After training, all participants were required to finish an object recognition task and rate the difficulty level (1 – easy, 3 – medium, 5 – very difficult) of the task.

During the recognition task, sound feedback was provided for users if their cursor collided with any object. At the beginning of each trial, users started from the center point of the 3D VE. In each trial, they needed to explore the VE and find the target object (sphere) in 90 seconds. After each trial, they evaluated the level perceived difficulty level of task. Their movement and performance were recorded during the task.

3 Results and Discussion

3.1 Task Performance

A summary of users' performance in each 3D VE is shown in Table 1. From Table 1, in the 3D VEs with symmetric layout, users moved faster and spent less time to complete the task and they had a higher completion ratio than the in the 3D VEs with random layout. Tasks in 3D VEs with a symmetric layout also had lower rating scores of difficulty levels. As the number of objects increases, generally their performance and perceived difficulty decreased. It is a natural result because as the number of objects increases, the chances to encounter non-target objects during motion increases.

Table 1. Performance of object recognition task

3D VE	VE1	VE2	VE3	VE4	VE5	VE6	Overall
Objects Number	4	8	12	4	8	12	1
Layout	Symmetr ic	Symmetr ic	Symmetr ic	Rando m	Rando m	Rando m	
Ave. Completion Time	49.911	42.375	56.073	64.641	55.51	84.193	55.451
Completion Ratio	0.833	1	1	0.667	0.667	0.333	0.75
Ave. Speed	0.063	0.288	0.639	0.084	0.082	0.094	0.208
Ave. Travel Distance	55.791	77.816	88.397	87.101	210.97 9	407.95 5	154.67 3
Ave. Difficulty Level	2	2.5	2.667	3.667	2.667	4.417	2.819

3.2 Effects of Objects Number and Layout

We applied a two-way ANOVA in order to test the effect of number of objects and layout type on performance and perceived difficulty. As shown in Table 2, users' completion time and task difficulty levels were significantly different between the two types of layout (random, symmetric). When the layout is uncertain by trial, they

should remember the layout every time, so it seems that it caused a significant burden for their cognitive process. The results show that the uncertainty of layout has a significant impact when users with visual impairments create the spatial mental map of a virtual environment.

On the other hand, the number of objects showed no significant effect on any performance variables. Even though the statistical effect was not significant, according to Table 1, performance was different for trials with differing numbers of objects. It seems that the differentiation of numbers of objects were not sufficiently large to show significant difference. In future studies, the effect of number of objects can be identified more clearly with more variation of this variable.

Table 2. The effect of number of objects and layout on users' performance

	Source	DF	F	P
Completion Time	Number of Objects	2	0.74	0.485
	Layout	1	10.55	0.003
	Interaction	2	2.14	0.135
Difficulty Levels	Number of Objects	2	0.33	0.723
	Layout	1	10.92	0.002
	Interaction	2	2.62	0.089

3.3 Observed Behavioral Characteristics

We analyzed users' behavioral pattern based on their movement profile. Their actual movements were recorded automatically using our haptic system with the frequency of thirty times per second. Then, the results were drawn using Matlab with the actual object used in the experiment (see Fig. 2). Based on six figures of their actual movement, we found some distinctive behavioral patterns, then analyzed the frequency of the found pattern based on the actual movement data to test whether the pattern can be supported or not.

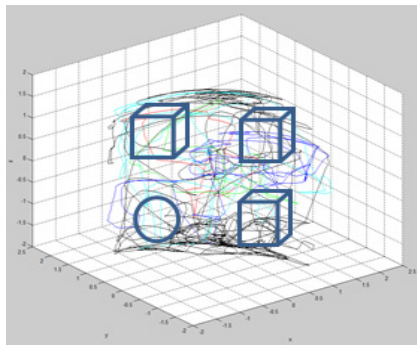


Fig. 2. User's actual movement data used for analysis

According to the analyzed pathway, behavioral patterns to identify the different objects were different. As shown in Fig 3. users tend to touch more frequently when they encounter a flat surface compared to round surfaces. The frequency analysis (number of collisions of one object during recognition) also support this tendency (average frequency of touching round surface: 3.34 (STD: 1.24); average frequency of touching flat surface: 7.76 (STD: 2.54)). This result indicates that recognition of a round surface is easier than polyhedrons with flat surfaces. We didn't test a polyhedron that has both round flat surfaces (e.g., cylinder), however, in future research if we use a cylinder, their behavioral strategy can be identified more clearly.

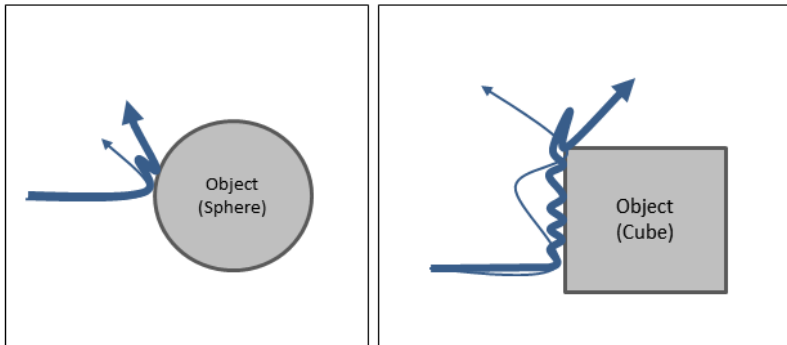


Fig. 3. Different behavioral patterns in recognition of different shapes (sphere and cube, thickness shows its relative frequency)

In exploration of free space between objects and objects, users also showed some typical behavioral characteristics. As shown in the left image of Fig. 4, most users preferred to reach the boundary of the environment first, rather than objects in the center of the space (total frequency of touching objects: 3674; total frequency of touching boundary; 1127). These results are in line with the results of Lahav and Moduser's study on behavioral characteristics of users with visual impairments in real environments (Lahav et al., 2003). According to the results, the users with visual impairments have an inclination to touch a boundary first as opposed to other areas in order to remember location of objects in reference to the location of the boundary. The results of our study confirmed that this inclination in real environments also applies to virtual environments.

Furthermore, participants showed a tendency to explore horizontally first then vertically (total travel distance in the horizontal direction: 3425cm, total travel distance in vertical direction: 2081cm). This result indicates a strategy to create spatial mental map. According to our results, they showed a tendency to use horizontal exploration for object recognition and spatial map creation, and vertical movement is used as a supplement (total frequency of touching objects with horizontal movement: 1536, total frequency of touching objects with vertical movement: 932). Based on this result, it can be explained that users with visual impairments have a tendency to create a spatial mental map first with horizontal relationships, then vertical relationships.

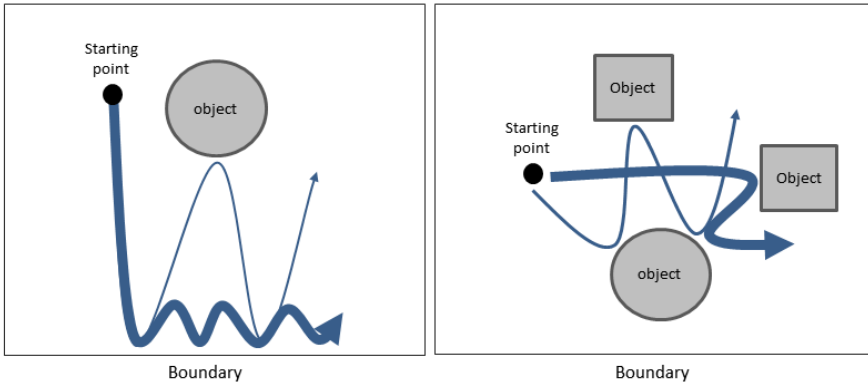


Fig. 4. Different strategies to explore the free space (thickness shows its relative frequency)

4 Conclusion

For users with visual impairments, haptically enhanced VEs can offer better benefits and opportunities by effectively providing users with information of the world using tactile information. In this study, we conducted an experiment in six haptically enhanced 3D VEs to study the behavioral characteristics of people with visual impairment. The layouts of 3D VEs were developed with different combination of two factors: different number of objects and different layouts. The results showed that these two factors can affect both user's manipulation behavior and performance. Also, the result of this study encourages the notion that haptics can provide shape and spatial information for users with visual impairment. These results can improve understanding of how users with visual impairment behave in a 3D VE. Additionally, this information may prove useful in developing guidelines for designers and researchers when developing assistive systems for users with visual impairment.

For future studies, quantitative models for behavior analysis such as State-Space Model (Jonsen, 2005) or Hidden Markov Model can be applied to get a better understanding of human movement in virtual environments. Also, more factors with a potential impact on behavior in 3D VEs, such as density and haptic feedback, could be examined in a future study.

References

1. Espinosa, M.A., Ungar, S., Ochaita, E., Blades, M., Spencer, C.: Comparing Methods for Introducing Blind and Visually Impaired People to Unfamiliar Urban Environments. *Journal of Environmental Psychology* 18, 277–287 (1998)
2. Fritz, J.P., Barner, K.E.: Design of a Haptic Data Visualization System for People with Visual Impairments. *IEEE Transactions on Rehabilitation Engineering* 7(3), 372–384 (1999)
3. Jonsen, I.D., Flemming, J.M., Myers, R.: Robust State-Space Modeling of Animal Movement Data. *Ecology* 86, 2874–2880 (2005)

4. Lahav, O., Mioduser, D.: A Blind Person's Cognitive Mapping of New Spaces Using a Haptic Virtual Environment. *Journal of Research in Special Educational Needs* 3(3), 172–177 (2003)
5. Li, Y., Johnson, S., Nam, C.: Haptically Enhanced User Interface to Support Science Learning of Visually Impaired. In: Jacko, J.A. (ed.) *Human-Computer Interaction, Part IV, HCII 2011*. LNCS, vol. 6764, pp. 68–76. Springer, Heidelberg (2011)
6. Roth, P., Petrucci, L.S., Assimacopoulos, A., Pun, T.: Audio-Haptic Internet Browser and Associated Tools for Blind and Visually Impaired Computer Users. In: *Workshop on Friendly Exchanging Through the Net*. (2000)
7. Särkelä, H., Takatalo, J., May, P., Laakso, M., Nyman, G.: The Movement Patterns and The Experiential Components of Virtual Environments. *International Journal of Human-Computer Studies* 67, 787–799 (2009)
8. Spence, I., Feng, J.: Video Games and Spatial Cognition. *Review of General Psychology* 14, 92–104 (2010)
9. Yu, W., Ramloll, R., Brewster, S.: Haptic Graphs for Blind Computer Users. In: Brewster, S., Murray-Smith, R. (eds.) *Haptic HCI 2000*. LNCS, vol. 2058, p. 41. Springer, Heidelberg (2001)

Part V
Graphical User Interfaces
and Visualisation

A Situation Awareness Assistant for Human Deep Space Exploration

Guy Andre Boy^{1,2} and Donald Platt¹

¹ Human Centered Design Institute, Florida Institute of Technology, 150 West University Blvd,
Melbourne, Florida 32901, U.S.A.

² NASA KSC, Mail Code IT-C1, Kennedy Space Center, Florida 32899, U.S.A.
{gboy, dplatt}@fit.edu

Abstract. This paper presents the development and testing of a Virtual Camera (VC) system to improve astronaut and mission operations situation awareness while exploring other planetary bodies. In this embodiment, the VC is implemented using a tablet-based computer system to navigate through interactive database application. It is claimed that the advanced interaction media capability of the VC can improve situation awareness as the distribution of human space exploration roles change in deep space exploration. The VC is being developed and tested for usability and capability to improve situation awareness. Work completed thus far as well as what is needed to complete the project will be described. Planned testing will also be described.

Keywords: Situation Awareness (SA), Augmented Reality, Human-Computer Interaction (HCI), Tablet Computing, Usability Testing, Space Exploration.

1 Introduction

The Virtual Camera (VC) for human deep space exploration is a virtual assistant that is being developed for use by astronauts and other associated mission operations personnel as they use human-piloted rovers to explore the surface of other planets [1]. The VC concept is based on incremental upgrading of a suboptimal 3-D geographical and geological database. It is an interactive window on the world as we know it at the time it is being used. Such an interactive window enables the user to maintain better situational awareness, to navigate and further explore space. It can be used for training, data analysis and augmentation of actual surface exploration.

The VC for human space exploration was originally described for use by astronauts navigating a surface exploration rover on a remote body such as the Moon [1]. Further analysis has indicated that such an interactive window to data in multiple dimensions has many other applications in domains such as aviation, medicine, control systems and finance. This paper presents a tablet version of the VC for deep space exploration using a scenario-based design approach. This work brings to human space exploration the use of advanced interaction techniques not previously used in spacecraft cockpit systems.

2 Motivation

There are significant challenges for astronauts while operating in the space environment, including increased levels of stress and workload, combined with decreased situation awareness. Astronauts and their exploration are evolving in an environment that is characterized by progressive discovery.

Human deep space missions will require increased roles for astronauts on-board as well as increased on-board automation. Previous work has determined that robots and humans need to work together in co-coordinated efforts [2]. It is envisioned that the VC can integrate data from both human-piloted and autonomous robotic explorers to improve the safe and efficient human exploration of a deep-space location (i.e., Moon, asteroid or Mars). On-board software and displays will integrate and manage all relevant systems as well as mission data. This information will include system health and status, caution and warning, traverse execution and mission timeline parameters.

There will be a need to make decisions based upon the collection and analysis of raw data to provide predictive information [3]. This information needs to be presented to crews in a way that enhances situation awareness. One way to do this is with new interactive environments such as 3-D tablet computer systems using advanced interaction techniques based upon accelerometer- and gesture-based inputs.

3 Related Work

Interactive cockpits, context-sensitive systems and remote agent knowledge representation are being developed for a number of domains including aviation [4], nuclear power plants [5] and even passenger car driving [6]. The VC will demonstrate these requirements in a system usable for deep space human exploration.

In early tests using a human-piloted rover in space-simulated environments it became clear that astronaut situation awareness needed to be improved [7]. Often when operating in environments such as the NASA Desert Research And Technology Studies (RATS) space exploration test bed, rover operators needs people outside the rover vehicle to guide for safe operation of the vehicle. This will not be possible for all rover or other space vehicle operations in deep space.

4 Design Goals

In deep space, situation awareness will be difficult due to the lack of input from the ground that astronauts are typically used to as well as the skill-retention concerns and environmental isolation from the environment under exploration.

The VC is playing the role of a remote agent for mission operations personnel and scientists. Functions that were previously done by human experts on the Earth for missions near Earth will be transferred to the on-board VC. With this transfer there will be emergent behaviors, which will be explored and identified through this research.

NASA is currently developing technology for use with deep space surface exploration. A centerpiece of this exploration effort is the Multi-Mission Space Exploration

Vehicle (MMSEV). This vehicle is being designed for use by astronauts on a variety of missions including Near-Earth Asteroid exploration. It is anticipated that the VC would support exploration conducted by this vehicle. Figure 1 shows a MMSEV concept that is based upon the Surface Exploration Vehicle that has been tested at the NASA Desert RATS analog test-bed. Part of the goal of designing a VC for space exploration is to afford the most natural and intuitive human-system interaction. A tablet computer interface provides a small, portable platform with a very common user interface that is gaining wide acceptance.



Fig. 1. Multi-Mission Space Exploration Vehicle (MMSEV) for Astronaut Deep Space Exploration (NASA)

A tablet allows gestures rotational interaction that astronauts have expressed that would be useful in the space environment. Two tablets are being used by Russian cosmonauts for recreation on the International Space Station [8], and by pilots for the use of electronic charts in commercial aircraft cockpits [9].

5 Work Completed

A human-centered design approach is being used to develop the tablet-based VC. The first step was to develop low-level prototypes to capture the basic design requirements for the VC. These were then presented to potential users during the fall of 2011 at the NASA Desert RATS test sessions. Horizontal prototypes were developed that captured the basic interaction requirements for the VC and were used to demonstrate an operational scenario. Currently, a complete vertical prototype is being developed with all of the interaction capability using a tablet PC device.

5.1 User Requirements Elicitation

The authors conducted a survey of potential users and stakeholders (astronauts, mission operations personnel, and scientists) for the VC system at the 2011 NASA

Desert RATS analog exploration testbed. The main questions during potential user interviews involved their background, general system uses, interface type and data display parameter formats desired.

User feedback indicated a device that provides real-time sensor knowledge display combined with multi-modal information display (vision and auditory for instance) would be useful. They also suggested that the VC should help in the navigation process and also display information about consumables such as power level, fuel levels and life support systems. This information was used to develop initial horizontal prototypes showing possible uses, screens and interactions for the VC tablet device.

Use cases for the VC were defined, and then a **horizontal prototype** was developed. This allows human-in-the-loop simulations to take place where the system can be tested with a variety of users. Other scenarios can also be illustrated in this way. Users can interact with actual system prototypes, which are being developed on the tablet-based computer system that will be used for the end product. This then allows the full functionality to be incorporated after initial user testing and feedback.

The horizontal prototype screens show the functionality and capabilities of the VC although the database interaction was not yet be possible. This horizontal prototyping allows astronaut, scientist and mission operations personnel to provide feedback as to the usefulness of the screens being developed.

Asteroid Itokawa has been modeled using a stereo lithography database format (STL) created by Gaskell [10] based on data from the Japanese asteroid mission Hayabusa. This data was used in the developed prototypes to simulate flying close to asteroid Itokawa and the resulting VC interactions.

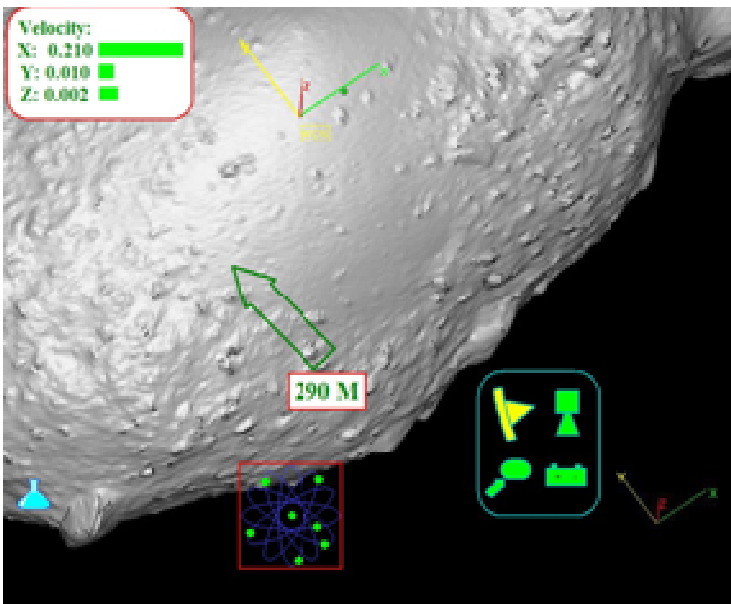


Fig. 2. The VC Horizontal Prototype Interface

The VC interface screen in Figure 2 shows the actual asteroid surface as projected by the STL file. This is how an astronaut would interface with the VC as they approached the asteroid in an MMSEV. The vehicle's velocity in X, Y and Z coordinated is shown in the upper left. Next to the actual velocity are bars of relative velocity. The color green represents a safe speed for the operating environment. If the velocity was slightly high, then the bar would turn yellow and if it was a dangerous level it would turn red. The velocity vector is simulated by the green arrow, which also points in the basic direction of travel.

An information cartouche in the lower right shows icons for communication link, thruster systems in use, electrical power and collision potential going clockwise from the upper left. Other information elements include a radiation sensor-warning icon in the lower middle indicating an on-board radiation sensor has detected radiation. In the lower left the blue beaker icon indicates an area of scientific interest.

In this example scenario the astronaut can determine to fly toward the beaker icon to attain a mission science goal or ignore the icon and fly in another direction. If another direction is chosen, other areas of interest could be displayed by icons. Sensor data is also compared to the current vehicle configuration and operational mode to determine possible safety concerns and out of nominal operation.

Figure 3 shows some of the safety information displays possible in the VC. In this case the astronaut is flying closer to the asteroid and is close to a mountain feature on the surface. Dust may be kicked up as he approaches and a dust warning has been triggered. The fuel level in the thruster triad currently active has reached a threshold that has triggered the rocket icon to turn red. Also the communications link with either the Earth or a nearby habitation system has deteriorated causing the communications (antenna) icon to turn yellow.

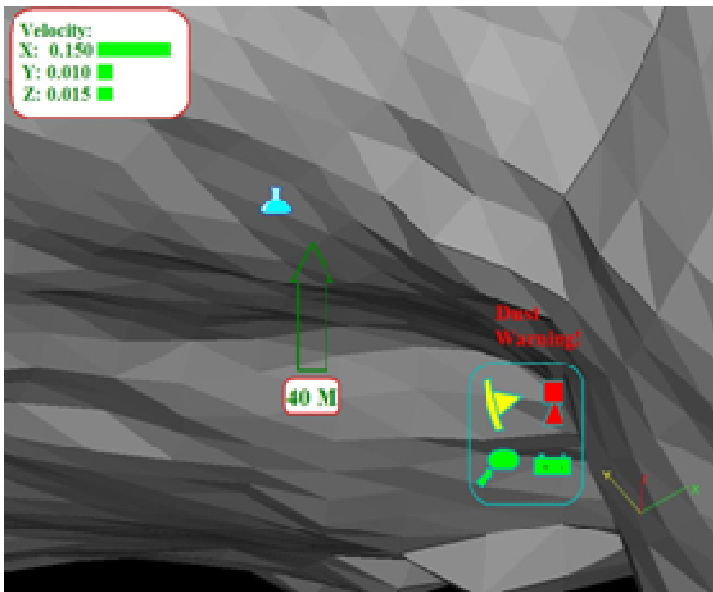


Fig. 3. VC Prototype Screen Showing Warnings

As the vehicle approaches the area of interest (in either a simulation mode or later in actual deep space exploration) the beaker icon changes to surface colors to indicate the exact area for exploration. This allows the astronaut to determine where to aim high resolution sensors, take high resolution images or to collect samples. Similarly mission operations personnel or scientists could “fly” mission scenarios and determine plans for exploration sorties at the actual asteroid. Annotations and other information screens can also be displayed as required when the exploration context calls for them. Examples could be radiation sensor alarms, loss of communication, and impending contact with the asteroid surface or a crater on the moon. Colors could represent either physical concentrations or an exploration priority scheme. These screens can be put together into a storyboard as the astronaut in the simulated MMSEV flies closer to and around the asteroid surface. This allows a movie to be created that simulates a mission scenario. Other types of information displays possibilities with the VC interaction include coloring the surface to correspond with certain areas of scientific or exploration interest. The elevation information would still be visible but additional information about areas that may potentially be explored could also be represented by different colors.

Emergent uses of the system and behaviors enabled or limited by the VC are discovered by the use of prototypes. A **vertical prototype** with the interaction functionality afforded by a tablet computer is now under development. This will allow complete testing of interactivity of the system. Figure 5 shows a screen from the VC vertical prototype using the Google Maps database. The information cartouche is displayed with the collision icon active if the user zooms in beyond a settable level. The direction arrow shows the direction of the accelerometer interaction sensor. The upper left shows where the VC application started as a central empty blue circle (in terms of geographic position) as well as a second red dot indicating the current view position in relation. Scale is also user settable. Scientific (beaker), resource (water drop) and user annotations (information) icons are also present.

6 Future Work

The final product of this work will be a tablet-based prototype VC device using the unique interaction capabilities of the tablet platform. These include portability, accelerometer-based 3-D interaction and the ability to be submerged into the 3-D environment. The interaction features such as icons, colors and annotations described are being incorporated.

This VC device will form the basis for evaluating the utility of the interactive database concept for improving situation awareness in the unique environment of deep-space human-piloted space exploration. Both mission planners and astronauts can conduct a simulated traverse using the VC, annotating areas of exploration interest or safety concern prior to conducting the actual traverse in the field. The tablet GPS location system can then be used to assist the astronauts as they follow through the traverse at a terrestrial space exploration analog site such as NASA Desert RATS.

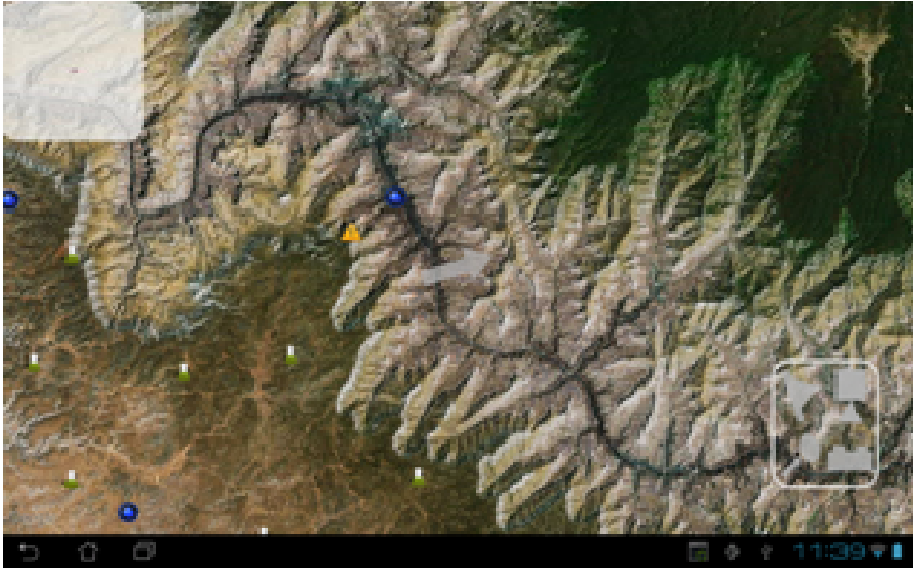


Fig. 4. The VC Vertical Prototype Interface Using Google Maps as the Database

7 Conclusion

The virtual camera for planetary exploration will assist in future exploration involving humans and robots. It is designed to provide improved situational awareness and augmented reality to define areas of interest and safety concern on remote planetary bodies. It will act as remote agents for mission planners and scientists capturing their knowledge and expertise as astronauts explore the solar system. It is claimed that the advanced interaction media capability of the VC can improve situation awareness as the distribution of human space exploration roles change in deep space exploration.

Human-centered-design techniques are being applied to the development of this tool involving all potential user groups from the beginning of the design process. Moving forward, the VC prototype will be finalized and further testing with domain experts will take place. Changes found in this testing will be incorporated into an updated VC and final evaluation will take place comparing this system to other interaction devices currently being developed. Recommendations will be made for future work to create a flight-ready VC capability. The advanced interaction possibilities with the VC will allow new types of displays and an immersive environment to assist astronauts to comprehend and project into the future situation awareness as their autonomy from ground control resources increases in deep space.

Acknowledgments. We would like to thank Ph.D. students, Kara Schmitt and Lucas Stephane for their advice and support. NASA Kennedy Space Center's Rebecca Mazzone and Michael Conroy have provided logistical support for this project.

Darrell Boyer and Matthew Long have provided software development support for this project. This research is also part of the French-American Project “Risk Management in Life Critical Systems” funded by the Partner University Fund.

References

1. Boy, G.A., et al.: The Virtual Camera: A Third Person View. In: Third International Conference on Applied Human Factors and Ergonomics (2010)
2. Deans, M.C., et al.: Field Testing Next-Generation Ground Data Systems for Future Missions. In: 42nd Lunar and Planetary Science Conference (2011)
3. Cummings, M., et al.: Past, present and future implications of human supervisory control in space missions. *Acta Astronautica* 62, 648–655 (2008)
4. Harper, R., Hughes, J.A., Shapiro, D.: Harmonious Working and CSCW: Computer Technology and Air Traffic Control. In: Bowers, J., Benford, S. (eds.) *Studies of Computer Supported Cooperative Work*. North-Holland, Amsterdam (1991)
5. US Nuclear Regulatory Commission: Technical Basis and Implementation Guidelines for A Technique for Human Event Analysis (ATHEANA), United States Government, Washington, D.C., NUREG-1624, Rev. 1 (2000)
6. Raja, A.K., et al.: *Power Plant Engineering*. New Age International (2006)
7. Boy, G.A.: The Virtual Camera Progress Report, report to NASA (2011)
8. Malik, T.: iPads Would Be Great in Space, Astronaut Says (2012), <http://www.space.com/14822-ipads-space-station-astronaut.html>
9. Paur, J.: FAA approves iPads for pilots’ electronic charts. *Wired*, Condé Nast, New York (February 28, 2011)
10. Gaskell, R., et al.: Landmark Navigation Studies and Target Characterization in the Hayabusa Encounter with Itokawa. In: AIAA 2006-6660, AAS/AIAA Astrodynamics Specialists Conf., Keystone, CO (2006)

My-World-in-My-Tablet: An Architecture for People with Physical Impairment^{*}

Mario Caruso¹, Febo Cincotti², Francesco Leotta¹, Massimo Mecella¹, Angela Riccio², Francesca Schettini¹, Luca Simone³, and Tiziana Catarci¹

¹ Sapienza Università di Roma

Dipartimento di Ingegneria Informatica, Automatica e Gestionale

{caruso,leotta,mecella,catarci}@diag.uniroma1.it

² IRCSS Fondazione Santa Lucia Italy

{f.cincotti,a.riccio,f.schettini}@hsantalucia.it

³ Sapienza Università di Roma, Dipartimento di Psicologia

luca.simione@uniroma1.it

Abstract. Mobile computing, coupled with advanced types of input interfaces, such as Brain Computer Interfaces (BCIs), and smart spaces can improve the quality of life of persons with disabilities. In this paper, we describe the architecture and the prototype of an assistive system, which allows users to express themselves and partially preserve their independence in controlling electrical devices at home. Even in absence of muscular functions, the proposed system would still allow the user some communication and control capabilities, by relying on non-invasive BCIs. Experiments show how the fully-software realization of the system guarantees effective use with BCIs.

Keywords: Brain Computer Interfaces (BCIs), tablet, home appliances, communication capabilities, software architecture.

1 Introduction

A cure for many neurodegenerative diseases is still unknown, yet advancements in life-supporting technologies and clinical practice allow a growing number of patients to survive longer. For instance, persons with Amyotrophic Lateral Sclerosis (ALS), undergo a degenerative process that lasts years, in which motor functions are progressively lost [7]; due to the heterogeneity of the disease (e.g., bulbar versus spinal forms), each patient experiences her own path of function deprivation; finally any chance of communication and action on the environment is lost; unless a fatal event occurs (e.g., a respiratory crisis) these individuals enter a locked-in state. While the advancement of life support technology and clinical practice can prolong the life of these subjects, it also extends the period in which her motor functions are very poor or even absent, leading to a state of

^{*} This work has been partly supported by the Italian Agency for Research on Amyotrophic Lateral Sclerosis (ARiSLA), through the project BrIndiSys - Brain-computer interface devices to support individual autonomy in locked-in individuals - <http://www.brindisys.it/>

complete dependence on the caregivers. As a consequence, social inclusion and quality of life of people with neurodegenerative diseases is decreasing, while the social cost for their assistance is increasing. Beside neurodegenerative diseases, other congenital or acquired deficits of the neuro-muscular system may lead to mild to severe limitations of mobility, motor skills, and speech.

In this paper we present the architecture and the prototype of an assistive system, referred to as My-World-in-My-Tablet (MWiMT for short in the following) suited for different inputs, fitting the residual abilities of the user, and aimed at preserving her communication ability at any stage of a progressing disease. The system allows the user to express herself and partially preserve her independence in controlling electrical devices at home. Even in absence of muscular functions, the proposed system still allows the user some communication and control capabilities, by relying on non-invasive Brain-Computer Interfaces (BCIs) [15]. In fact, by relying on modulation of brain activity voluntarily induced by the user, and detected by processing her electroencephalogram (EEG), BCI research has shown in the past decade the possibility of a communication even in absence of any muscular contraction.

MWiMT is based on a tablet device, and consists of two main software components: AUXILIHOME, which provides basic communication tools and a flexible access to home automation appliances, and FLEXYGLASS, which allows operating different mainstream applications using a common interface supporting different kind of aids. Its design allows an early adoption of the aid, when the user can still operate it by means of conventional interfaces (e.g., a manual or automatic scan button), and can be re-configured whenever the user, due to her decay of motor abilities, feels no more able to operate it.

2 Preliminaries

The term *assistive technology* (AT) originally included all kinds of accessible, adaptive and rehabilitative devices addressed to people with disability, aimed at improving their activities and participation and thus their quality of life. Nowadays the term significantly changed its meaning, including, in addition, a wide variety of software solutions which replace, and in some case improve, the features originally provided by specific devices. *Augmented and alternative communication* (AAC) is a classical AT application aiming at compensating for severe speech-language impairments in the expression or comprehension of spoken or written language. Software packages for communication, supporting different inputs, running on a common personal computer are available; they can simulate communication boards (both alphabetical or symbolic), reproduce virtual keyboards, and can be equipped with word prediction systems and vocal output, optionally giving access to the internet. Home automation (*domotics*) represents another promising AT application area [12,8,3].

Input devices for AT systems can be classified into two main categories, namely *pointing devices* (e.g. trackballs, joysticks, touch screens, trackers) and *switches* (or more generally binary input devices, used in combination with automatic scan or step scan systems). Beside these classical input methods, last

years have seen a growing interest in *brain computer interfaces* (BCIs) and more generally in biosignal based interfaces. BCI is intended as a mean for providing severely physically impaired people (locked-in subjects) a direct communication channel between the brain and an external electronic device. In particular, many studies have been conducted with so called non-invasive BCIs in order to translate electroencephalographic (EEG) activity or other electrophysiological measures of brain function into commands [2]. A set of different EEG features which can be translated into control signals together with the needed processing steps are described in [15]. A BCI translates these features into control signals either continuous or discrete in time.

While BCI performance improved over the years, required hardware and software became cheaper and simpler to use giving the chance of bringing BCI directly at home without the continuous support of a specialized technician. Nowadays, portable EEG amplifiers can be accessed by a computer using standard interfaces (e.g. RS232, USB, Bluetooth). Additionally, a set of established software BCI platforms is freely available for real-time EEG signal processing (e.g., BCI2000, OpenVibe). However, currently, a few demonstrations of BCI as possible assistive product have been given and few cases are reported in the literature of motor disabled users that can access to communication and environmental control through a BCI [14,2].

3 The System

The goal of our work is the definition of a modular user-centric platform, depicted in Figure 1, in which an off-the-shelf tablet is used as a generic and extensible integration container for a set of different technologies.

Input methods are split into *hardware-based* and *biosignal-based* ones. Different input methods allow for different information transfer rates from the user to the controlled devices. A user usually chooses the input method which, by taking into account her abilities, degree of impairment and required effort, offers the higher information transfer rate. A key advantage of such an architecture is the transparency of the input method with respect to the controlled application.

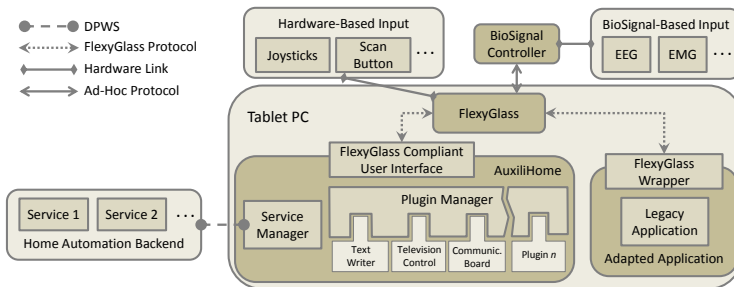


Fig. 1. My-World-in-My-Tablet architecture

Using a biosignal as input method is only possible using a biosignal controller, which is the component (hardware or software) devoted to the translation of the biosignal into a *control signal*. This work focuses on BCI as biosignal-based input method so, in the following, we will refer to this component as the BCI controller. This component may run directly on the tablet as well as on a different computer connected to the acquisition hardware. Its output may provide either a discrete control signal or a continuous one. At the actual stage of development, our prototype relies on BCI2000 as a BCI controller providing a discrete control signal based on P300 feature (see Section 3.2) of EEG.

The employed *home automation back-end* [3] hides all the home automation functionalities behind well-defined software services; it has been designed to be enough versatile to be application agnostic, allowing to detect, use and compose every kind of device on a *semantic* base.

The tablet represents the way a user with impairments is able to communicate her needs and ideas and to control the domestic environment using the available input methods. This mobile device runs several different applications, most of which have not been designed to interact with special input aids. The most innovative component of our architecture, namely FLEXYGLASS, has been devised as a way to provide a standard interaction method with installed software, despite the variety of available input methods for the user. An example of such a software is given by the AUXILIHOME component, which gives access to a set of application plugins and to the services provided by the home through an adaptive and extensible GUI.

Personalized dynamic accessibility [6,4] aims at achieving a more effective user interaction by making the software adaptive with respect to user's needs and abilities changing over time. This is obtained through the *customizability* and the *dynamic adaptivity* of the user interface. Our work pursues the same purposes providing respectively (i) the possibility to configure and personalize the applications composing AUXILIHOME (see Section 3.1) and (ii) the possibility of using several different input devices and modalities thanks to the novel FLEXYGLASS subsystem (see Section 3.2).

3.1 AuxiliHome

The graphical interface of AUXILIHOME consists of several screens, each of them based on a grid layout into which graphical components can be placed. The number of rows and columns of the grid can be configured and determines the minimum size of a graphical component. Such a grid layout is easily described in a declarative manner, through a specific XML document, to be offered by services and external applications.

Grid positions are occupied by buttons (see Figure 2) which, assuming different sizes, allow to strengthen the visibility of specific objects according to specific user's needs. A minimalistic interface has been targeted thus avoiding irrelevant or rarely needed information. The default look of the graphical user interface was designed aiming at complying with usability and accessibility guidelines [9]. Buttons come with a matte black background and white icons and labels. Only

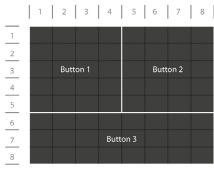


Fig. 2. UI elements

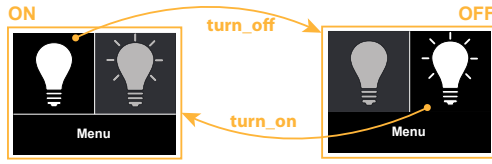


Fig. 3. Light service model

these two colors were used to code information in the whole interface and they maximize contrast ratio. The big contrast and the uniform background improve to screen clarity and readability.

AUXILHOME is a collection of applications that implement functionalities needed by users with disability; such applications are accessible through the graphical interface. The level and type of disability vary from user to user and may evolve over time, thus, the organization of applications should be as dynamic and flexible as possible. Applications can be added or removed easily from the system whenever necessary, must be configurable and, in order to provide maximum expandability of the system, can be developed and deployed by third parties. Applications can be either *tightly* or *loosely* coupled to the graphical interface and to the tablet.

Generally speaking, we can say that all the home automation applications can be considered loosely coupled to the tablet, i.e., they run outside the tablet since they belong to a specific home environment and installation. Both sensors and actuators in the home make their functionalities available according to a *service oriented architecture* (SOA) approach [3], which employees Web services as a way to face the heterogeneity of device’s specific protocols. According to a rich service model, a Web service consists not only of the service interface specification, but also of its conversational description and of the graphical widgets (i.e., icons) needed when presenting the operations in the user interface. *Light control* for example (see Figure 3) allows the user to turn on or off a light. In this case the behavior of the application and of the UI can be easily modeled with a descriptive approach. Similar services related to armchairs, beds, alarm bell and doors have been employed during validation.

Among the currently available technologies for implementing Web Services, we chose Devices Profile for Web services – DPWS. By relying on DPWS, devices are discovered as soon as they join the local area network and the appropriate application is dynamically loaded on the tablet whilst, on the other hand, the asynchronous *event driven architecture* allows the graphical user interface to immediately reflect changes in the state of each device.

Tightly coupled applications are implemented as *plugins*. From this point of view, AUXILHOME behaves as a host application, which provides common functionalities that a plugin can use (e.g., the speech-engine). Plugins must implement a simple interface consisting of some methods that are executed during their life-cycle. This solution grants maximum expandability to the system, allowing third party applications to be easily developed and deployed only focusing

on the core functionalities, without worrying about the graphical interface or the input mechanism that are under the responsibility of the AUXILIHOME container.

Two applications/plugins for communication purpose are already available. The first one is a speech synthesizer for frequently used sentences. The list of sentences is fully customizable by the user or by the caregiver. The second one consists of a virtual keyboard provided with a word completion system and a speech synthesizer; it additionally shows a flipped copy of the inserted text at the top of the screen to make possible face-to-face conversations. Moreover plugins for infrared controlled devices such as TVs and DVD players are provided.

3.2 FlexyGlass and Adapters

The FLEXYGLASS component is an independent software module which makes the employed input method totally transparent to the controlled application. The basic idea (see Figure 4) behind FLEXYGLASS, is to over impose a transparent pane to the controlled application UI (using a topmost window with a transparent background); such a transparent pane contains a set of virtual controls which inherit size and position from the controlled application real controls (buttons, links or focusable objects) and which act in principle as proxies: if a virtual control is selected using the input method chosen by the user, the corresponding real control is triggered.

The described approach requires a direct connection between a controlled application and the FLEXYGLASS, and a communication protocol allowing FLEXYGLASS to (i) request the list of currently available controls to the controlled application in order to update virtual controls, and (ii) to communicate the last selection triggering the execution of a real command (the controlled application will acknowledge the completed execution of the command).

The FLEXYGLASS component is intended to support different kinds of input methods, each one requiring a different kind of interaction. A scan button, for example, can be used for either (i) *manual scan*, which uses short pressures to move the focus between controls and longer ones to trigger a command, or (ii) *automatic scan*, which automatically move the focus and interprets each pressure

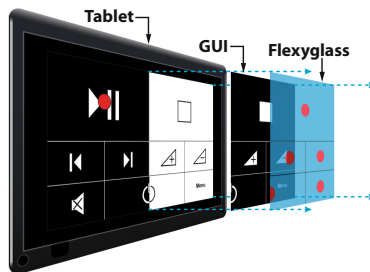


Fig. 4. A layered perspective of FLEXYGLASS

as a trigger. User who are able to move a mouse, but not to click, may use a *dwell mouse*, which automatically click after a predefined dwell time.

Currently, FLEXYGLASS supports P300 as the BCI-based input method. The P300 potential is a large and positive deflection of the EEG activity which reaches a maximum of amplitude over the centro-parietal scalp areas between 250 and 400 ms after a relevant stimulus (*target* stimulus), presented within a sequence of frequent irrelevant stimuli (*non-target* stimuli), is recognized [10]. The *speller* paradigm [5] is based on a n by m selection matrix divided into stimulation classes, one for each row and column composed by a set of symbols stimulated together. A *trial* is a stimulation sequence made up by a fixed number of consecutive shuffling of the whole set of stimulation classes. A specific symbol lays at the intersection between a row and a column, thus, if we suppose to have k repetitions, a user concentrating on a symbol will see it flashing $2k$ times (each followed by a P300 potential) within a train of $2k(n + m)$ stimuli. At the end of a trial, a score is assigned to each stimulation class; the system then selects the row and column with the highest score and returns the correspondent symbol.

The usage of P300 as an overlaid stimulation interface was first introduced in [11]. Here a P300 overlaid interface was used to control the commercial assistive technology application suite QualiWorld. FLEXYGLASS makes that idea more general, by allowing to use coherent graphics to control generic applications with different kinds of inputs ranging from BCIs to hardware switches. Using a BCI input requires a BCI controller (BCI2000 in our case) which has to be connected using an ad-hoc protocol. Creating an overlaid interface raises the problem of how the classical matrix layout of the P300 speller may be adapted to a more general layout. Before a trial begins, FLEXYGLASS analyses the controls available on the controlled application, chooses the minimum matrix size with enough space to define a one-to-one association between controls and matrix positions.

Controlling an application using P300 requires a continuous attention to the stimulation in order to avoid incorrect selection bringing the controlled application into an unwanted state. A user might desire to pause the system because she is either tired or occupied in some other tasks. FLEXYGLASS proposes a mechanism consisting in stimulating over the window shown in Figure 5.

Controls has to be highlighted in all previous input methods. By knowing the position of the available controls of the controlled applications, FLEXYGLASS is able to move the focus over a control by drawing over the layered window.

FLEXYGLASS have been designed to be easily extensible with different high-light graphics. Figure 6 shows the currently available stimulations, the *dot* [11]

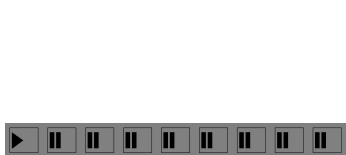


Fig. 5. Pause for P300 input

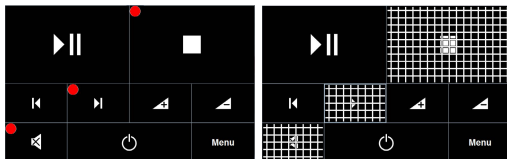


Fig. 6. Stimulation over the DVD remote control

and the *grid* [13] ones, which have been proved to be very effective with P300-based BCIs. The FLEXYGLASS graphical configuration utility allows for customizing each aspect of the available stimulations (colors, sizes, etc.).

Applications which are compliant with FLEXYGLASS are automatically discovered through a `hello` protocol. There exist two kinds of application: those which are natively built to be controlled via FLEXYGLASS (it is the case of AUXILIHOME) and those (the vast majority), which we refer to as *legacy* applications, which require the development of a *wrapper* in order to be controlled. In the case of a browser like Firefox or Chrome, a wrapper may be easily implemented as an extension.

4 Validation and Experiments

Usability, reliability and learnability of MWiMT have been specifically assessed in an experimental protocol, which includes: *(i)* communication task: users were requested to spell predefined sentences; *(ii)* environmental control: users were requested to perform some actions on the smart home. In order to provide a reference level to estimate accuracy and reliability, users were also involved in a BCI session with a widely validated P300-based Brain Computer Interface (the BCI2000 built-in speller). User satisfaction was measured with a visual analogue scale (VAS) at the end of each condition. Users were asked to rate their “overall satisfaction”, drawing a vertical bar on a line where number 0 indicated that they were “not satisfied at all”, whereas number 10 meant that they were “absolutely satisfied”. At the end of each session, users were also administered with the System Usability Scale (SUS), assessing the perceived satisfaction and usability with a score ranging between 0 and 100.

Table 1 reports the results of the experimentation over three end-users. The Amyotrophic Lateral Sclerosis Functional Rating Scale - revised (ALSFRS-r) [1] is an instrument for evaluating the functional status of patients with ALS. It can be used to monitor functional change in a patient over time. Score ranges from 0 to 48 and the higher the score the more function is retained. All users were able to complete the proposed tasks; they reached on average the 95% classification accuracy with MWiMT, conversely, the accuracy achieved with the classical

Table 1. Results of the experimentation

<i>User</i>	<i>ALSfrs-r</i>	<i>Device</i>	<i>Task</i>	<i>Accuracy</i>	<i>Satisfaction</i>	<i>SUS</i>
User1 (F) Age 75	38	P300 Speller	Communication	100%	9.7	42.5
		Prototype	Communication	100%		
			Environment Control	89%	10	70
Strong dysarthria, no experience with ATs						
User2 (M) Age 56	37	P300 Speller	Communication	100%	10	82.5
		Prototype	Communication	100%		
			Environment Control	100%	8.3	60
Slight dysarthria, motor impairment upper limbs, no experience with ATs						
User3 (M) Age 59	9	P300 Speller	Communication	95%	9.8	77.5
		Prototype	Communication	89%		
			Environment Control	90%	10	95
Severe motor disability. Residual movements: head, eyes, one finger of both the hands (very weak movements). Experience with ATs: communicator (used slowly with the finger)						

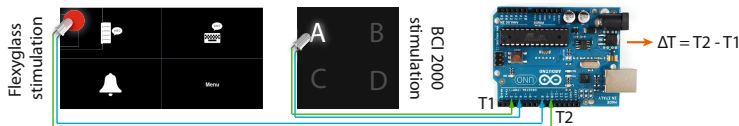


Fig. 7. The hardware platform employed for the feasibility test

P300 Speller ranged from 95% to 100% . Users expressed an high satisfaction with both the P300 Speller (values ranging between 9.7 and 10) and MWiMT (values between 8.3 and 10). The perceived usability, measured by means of the SUS, was on average 67.5 for the P300 Speller and 75 for MWiMT.

Besides the validation with the users, we performed some tests in order to analyze the visualization delay of the stimuli over the FLEXYGLASS layered window. The BCI2000 built-in speller is directly connected to the sequence generation allowing for the best delay between the visualization request and the effective onset of the stimulus on the screen. In our case, the stimulation sequence is transmitted using an interprocess communication channel introducing an unpredictable transmission delays. This delay is the time difference between the instant a stimulus is shown on the BCI2000 interface (taken as a “gold standard”), and that one the correspondent stimulus is shown over the FLEXYGLASS. A hardware test bed (Figure 7), consisting of 2 photo-transistors connected to the analog inputs of an Arduino One board, has been designed at this aim. The two photo-transistors are placed directly over the monitor screen; one over an element of the BCI2000 speller and the other one over the corresponding control of the FLEXYGLASS transparent window. The board is in charge of detecting light flashes onsets on both transistors and calculating the time difference.

Two experiments have been performed and results are shown in Figure 8. In both cases a sequence of 20 trials has been performed with 10 repetition of stimulus classes. During the first experiment, the BCI2000 ran directly on the tablet together with FLEXYGLASS and AUXILIHOME. We can see how the third quartile of delay measurements are below 8.2 ms while the half of the measurements are comprised between 4.9 (first quartile) ms and 8.2 ms. During the second experiment, BCI2000 ran on a separate machine (an off-the-shelf laptop) connected via wired network to the tablet. The third quartile is now significantly higher then before (about 16 ms) while the maximum is of 51 ms.

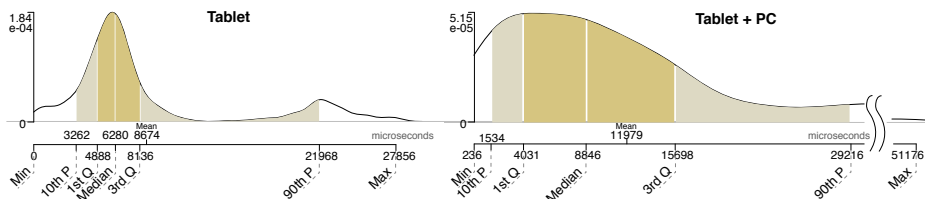


Fig. 8. Density function of measured delays (P - Percentile, Q - Quartile)

Despite the fact that our tests showed a measurable delay in the FLEXYGLASS stimulation, this does not impact on the P300 BCI performances. The P300 waveform shows 1 or 2 order of magnitude slower time constants, so recognition will not be affected by such a small relative time shift of the waveform.

5 Concluding Remarks

This work presented a general architecture, and the related prototype, based on a tablet device, for allowing physically impaired people to interact with the surrounding in a fully automated home environment. The proposed approach offers extensibility in terms of provided applications and input methods, and adaptivity in terms of automatic adaptation to the home automation system.

This is possible through a specific software architecture, whose pillars are the wrapping of home appliances as Web services, a plugin approach for incorporating specifically designed applications, and the FLEXYGLASS component for incorporating legacy applications and allowing specific input methods requiring stimulation (such as BCIs). The FLEXYGLASS may be used, in principle, in combination with every kind of application (provided with a specific adapter).

We have validated our approach with real users and demonstrated, through some tests, that introducing FLEXYGLASS as overlay over existing applications does not degrade performances of the input methods, even in the case of BCIs.

A note should be reported about the possibility of using whichever application in conjunction with FLEXYGLASS. This is not completely true if the selected input method is P300 BCI. In fact P300 classification is very influenced by stimulus distribution over the screen, suffering in particular of stimulation very close one to each other. A possible way of addressing this issue could be the possibility for the FLEXYGLASS to automatically analyse the set of provided controls and reorganizing it using call out and subsets definition. Additionally while FLEXYGLASS is able to automatically detect compliant application (native or legacy), no accessible interface is currently provided to select among these applications or switch from an application to another one. A future enhancement is about the menu structure of AUXILHOME, which is currently fixed. Unfortunately, some aids provide a low information transfer rate which can potentially make a single selection expensive. Future version of AUXILHOME could provide a menu structure which dynamically evolve following user favourite selections (learnt over past executions).

References

1. Cedarbaum, J., Stambler, N., Malta, E., Fuller, C., Hilt, D., Thurmond, B., Nakanishi, A.: The alsfrs-r: a revised als functional rating scale that incorporates assessments of respiratory function. *Journal of the Neurological Sciences* 169(1), 13–21 (1999)
2. Cincotti, F., Mattia, D., Aloise, F., Bufalari, S., Schalk, G., Oriolo, G., Cherubini, A., Marciani, M., Babiloni, F.: Non-invasive brain–computer interface system: towards its application as assistive technology. *Brain Research Bull* 75(6), 796 (2008)

3. Di Ciccio, C., Mecella, M., Caruso, M., Forte, V., Iacomussi, E., Rasch, K., Querzoni, L., Santucci, G., Tino, G.: The homes of tomorrow: service composition and advanced user interfaces. *ICST Trans. Ambient Systems* 11(10-12) (2011)
4. Edwards, W.K., Bellotti, V., Dey, A.K., Newman, M.W.: The challenges of user-centered design and evaluation for infrastructure. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 297–304 (2003)
5. Farwell, L., Donchin, E.: Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and Clinical Neurophysiology* 70(6), 510–523 (1988)
6. Gajos, K.Z., Hurst, A., Findlater, L.: Personalized dynamic accessibility. *Interactions* 19(2), 69–73 (2012)
7. Hardiman, O., van den Berg, L.H., Kiernan, M.C.: Clinical diagnosis and management of amyotrophic lateral sclerosis. *Nature Reviews Neurology* 7(11), 639–649 (2011)
8. Helal, S., Mann, W.C., El-Zabadani, H., King, J., Kaddoura, Y., Jansen, E.: The Gator Tech smart house: a programmable pervasive space. *IEEE Computer* 38(3), 50–60 (2005)
9. Nielsen, J.: Ten usability heuristics (2002), http://www.useit.com/papers/heuristic/heuristic_list.html
10. Polich, J., Kok, A.: Cognitive and biological determinants of p300: an integrative review. *Biological Psychology* 41(2), 103–146 (1995)
11. Riccio, A., Leotta, F., Bianchi, L., Aloise, F., Zickler, C., Hoogerwerf, E.-J., Kuebler, A., Mattia, D., Cincotti, F.: Workload measurement in a communication application operated through a P300-based Brain-Computer Interface. *Journal of Neural Engineering* 8(2) (2011)
12. Roberts, J.: Pervasive health management and health management utilizing pervasive technologies: Synergy and issues. *Univ. Computer Science* 12(1), 6–14 (2006)
13. Tangermann, M., Schreuder, M., Dahne, S., Hohne, J., Regler, S., Ramsay, A., Quek, M., Williamson, J., Murray-Smith, R.: Optimized Stimulation Events for a Visual ERP BCI. *Int'l. Journal of Bioelectromagnetism* 13(3), 119–120 (2011)
14. Vaughan, T., McFarland, D., Schalk, G., Sarnacki, W., Krusienski, D., Sellers, E., Wolpaw, J.: The wadsworth bci research and development program: at home with bci. *IEEE Transactions on Neural Systems and Rehab. Eng.* 14(2), 229–233 (2006)
15. Wolpaw, J.R., Birbaumer, N., McFarland, D.J., Pfurtscheller, G., Vaughan, T.M.: Brain-Computer Interfaces for communication and control. *Clinical Neurophysiology* 113(6), 767–791 (2002)

AHPM as a Proposal to Improve Interaction with Air Traffic Controllers

Leonardo L.B.V. Cruciol and Li Weigang

University of Brasilia, Brasilia, DF, Brazil
leocruciol@gmail.com, weigang@cic.unb.br

Abstract. Air Traffic Management (ATM) involves a complex decision-making process that involves several entities as short time to analyze risk situations and many attributes to verify before take an action. So, Decision Support System (DSS) is a great way to air traffic controllers achieve better results in their work. A well implemented DSS must provide a simple Human-Computer Interaction (HCI) to achieve great results. Even a system can provide all functionalities for a specialist, it must achieve his expectations and results by other requirements, i.e., maybe a right answer with delay or hard to find will become a wrong or unnecessary answer. The proposed approach by Air Holding Problem Module (AHPM) has a sub module responsible for forecasting airspace scenarios and another responsible to support decision-making process by an interaction with air traffic controller. Thus, it is possible that air traffic controller interacts with the system and carries out his activities faster and more informed by a simple screen which contains knowledge necessary. The AHPM achieved a great human-computer interaction level because the interaction is very simple and all mandatory information to do great analysis is presented in a same screen by a clean and objective organization.

Keywords: Human-Computer Interaction, Decision Support System, Air Traffic Management, Artificial Intelligence.

1 Introduction

Air Traffic Flow Management (ATFM) is considered a complex decision-making process that involves several entities as: aircraft and human being safety; short time to analyze risk situations; many attributes to verify, analyze and decide about the best group of actions to improve the air traffic flow. There are so many factors related to weather conditions, aircraft operational limitations and human capability to act in a short time interval under high pressure.

Human beings and machines are complementary in several aspects. The power of a taken decision by a human being in areas such as intuition, conceptualization and creativity are the weak points of a working machine. Human weakness, on the other hand, consists in aspects that a computer is accurate to achieve such as speed, parallelism, accuracy and the persistent storage of almost unlimited detailed information. So, a well implemented decision support system could help

air traffic controllers to take the best actions by a strong Human-Computer Interaction (HCI) that will use the best points of each one.

Moreover, the system must provide a simple interaction to achieve better results. Even a system can provide all functionalities for a specialist, it must achieve his expectations and results by other requirements, i.e., maybe a right answer with delay or hard to find will become a wrong or unnecessary answer. Several factors are essential to reach a great HCI level in ATFM domain, such as key features available by a click, an integrated knowledge base presented in a main screen, alerts graphics for easy perception when status had been changed, interaction with other features without get out of main control screen, and others.

Air Holding Problem Module (AHPM) has four sub modules. Among them, there are sub modules responsible for forecasting scenarios and interaction with specialist. Thus, it is possible that air traffic controller interacts with the system and carries out his activities faster and more informed by a simple screen which contains knowledge necessary.

The Forecast Scenarios Module is responsible for assessing the current scenario, verify possible risk situations and its solutions in accordance with system knowledge. However, as important as the whole process of prediction scenarios is to present clearly and quickly the system knowledge for the air traffic controller detects possible problems and acts quickly. In the ATFM domain is indispensable that actions are taken with great knowledge and in the shortest time possible. In a real-time problem, the best solution for the time T_n probably will not work at a future time T_{n+1} .

This paper presents the decision support system AHPM developed to act on ATFM scenario in Brazil. So, it was modeled considering the reality of the country and its air traffic controllers to achieve more effective results. The paper is organized in the following manner. In section 2, there is brief review of related concepts about Decision Support System and Human-Computer Interaction. Section 3 presents the environment of ATFM, which AHPM acts to support daily tasks by interaction with air traffic controllers. Section 4 presents the decision support system AHPM. Section 5 concludes the paper and proposes the direction of future research.

2 Decision Support System

The decision support systems can be defined as systems that support decision-making process by providing relevant information, suggestions and predictions, which are based on current information to provide a vision of the future, according as some actions are taken in the present.

The business processes that will be automated by a system must be chosen carefully. Specially about control activities; conflict detection and analysis, research and planning execution. Decision Support System (DSS) allows using data and models related to an area of interest to solve problems, semi-structured and unstructured, with which they are confronted to achieve a better system (Beulens et al., 1988; Bayen et al., 2005).

A DSS allows working with problems of a decision-making which the proportion overtakes the normal rational capacity or exceeds temporal and financial means available. The air traffic controller reports his difficulties to take actions with minimum impact in the future, so it can be represented in a system with management and control of existing organization knowledge.

According Agogino et al. (2009), it is essential that systems to support air transportation can be prepared to provide a flexible and automated management to meet requirements inherent in this kind of management. These systems are included in a new generation, which should be prepared to meet this demand.

Among the approaches that are presented in the literature, it is possible to classify a decision support system in four different ways of operation:

1. Without autonomy: the system displays information and the expert must check in several points what is useful, or not, for every situation.
2. Full Autonomy: the system, based on previously acquired knowledge, analyzes each situation and take its decisions.
3. Semiautomatic (more automatic): the system has enough intelligence to assess different situations and as situation decide itself. In other situations, the expert decides what should be done.
4. Semiautomatic (more human): the system has enough intelligence to analyze situations and make suggestions for solutions to the specialist, which will decide what should be done.

The approach of this research follows the fourth way presented. It will always leave the decision-making power with the air traffic controller. However, it will analyze situations and make suggestions to be taken to the specialist. This choice was made because of concern about safety of the airspace. So the air traffic controller will have the information generated by the system but with full autonomy to choose the AHPM suggestion or a new action according to your experience. When specialist decides for new actions, system will learn and suggest these actions for similar scenarios in the future.

Thus, the improvement of human-computer interaction becomes more important because must provide a knowledge base in the best way, so air traffic controller can carry out his activities achieving the best benefits of DSS. The system will be a major provider of knowledge and its interaction with the specialist will make the level of success for ATFM.

2.1 Human-Computer Interaction

The Human-Computer Interaction (HCI) field is responsible to improve how human being interacts with computer systems. There are so many researches to improve HCI covering software engineering, system usability, new approaches to interaction, multimedia technology, knowledge architectures, system design, cognitive computing and others.

Important as the adoption of techniques to improve HCI is continually checking the degree of satisfaction of each user, too. Thus, it will happen a continual

improvement process in the interaction in order to achieve a natural interaction between both its. Additionally, the system must be self-adapting as a specialist that use. So, it is possible to make a well experience for all system users in a same level (Leadbetter et al., 2000).

This improvement process should not only check the system, but the business processes which DSS is supporting too. It is possible to remove some complex spots that hinder human-computer interaction and add these points inside of the system, reaching more gains for the air traffic controller such as: time, handled complexity, reduce the impact of actions, take actions more effective and others.

It is necessary to analyze how the process is being automated by DSS and evaluate the negative impact that may be generated, such as semi complete automation generating omission in operation of DSS and make obscure the decision-making process to air traffic controller can decide to accept a suggestion of restrictive measure.

In domains more complex such as ATFM, this may forbid full adherence to the system by the system user because it is unknown what is happening inside the software. The objectives of DSS must walk together to aid air traffic controller instead of hiding everything that is being done, so this approach follows the standard semiautomatic (more human), i.e., making it clear for the air traffic controller to choose his decision (Yoshikawa , 2003; Grudin, 2009).

3 Improving Interaction in ATFM

Air Traffic Flow Management focuses on the supply of information to maintain the traffic flow with safety and reduced impact on airspace scenarios that are necessary to take unexpected measures. The ATFM environment can be organized into three phases: strategic, operational and tactical.

This paper focuses on ATFM tactical level because it is the period which aircraft is in flight. This level consists on tactical decision making covering the period from two hours before the flight until the aircraft arrives at its destination. During this phase increases the problem level because the occurrence of problem and its solution happens on real time. This factor also needs to be focused on HCI to improve the solutions for risk situations in ATFM in real-time environment.

The main problem to be resolved in this work is the Air Holding Problem (AHP), which occurs when aircraft in flight route needs to wait on the air for a particular reason, such as closed airport, hard meteorology conditions, terrorist acts, and others. These situations may impact in other areas of the far Terminal Maneuvering Area (TMA), place where an airport is situated. These impacts can be spread throughout the air traffic flow arriving at one departure airport, and thus preventing an aircraft take off.

The Brazilian airspace covers the entire territory of the country, including part of the Atlantic Ocean. In the airspace of Brazil there are five Flight Information Region (FIR): FIR - Amazon (north); FIR - Recife (northeast); FIR - Brasilia (midwest); FIR - Curitiba (south) and FIR - Atlantic (Atlantic Ocean coast). The FIR's are subdivided into sectors of control, to improve the management

activities and obtains better control. Currently, there are 46 control sectors, 14 in FIR - Amazon, 8 in FIR - Recife, 12 in FIR - Braslia, 10 in FIR - FIR in Curitiba and 2 - Atlantic.

These sectors are under supervision of air traffic controllers that are in an Area Control Center (ACC), which is responsible for a specific FIR. In this context, it is possible understand the complexity of management activities and why there are subdivisions to manage so many factors, e.g., the number of aircraft per sector influences directly the management complexity, i.e., the more aircraft flying in the same sector, more security risks involved in the ATM.

Given this context and in order to support ATFM was proposed Air Holding Problem Module (AHPM) as a new approach interaction with air traffic controllers to provide support to decision-making process and improve results on AHP. Currently, air traffic controllers use a control system in standard monochrome and basic screens, i.e., there are basically one screen to the standard radar display which they monitor the traffic flow, detecting possible risk situations and their solutions.

The problem with this model is that system is limited in the presentation of information. All hard work needs to be done by an expert in a short time and basically with a radar screen to survey the necessary information to take his decision. This new approach was shaped for air traffic controller could have more benefits of DSS such as providing knowledge on a screen instead of only some data, add on a screen the current traffic, possible solutions and their impacts, reduce the level of tiredness of their eyes through the visual alerts which reducing the need for high concentration on the screen, and others.

4 AHPM

The Air Holding Problem Module system was developed using two techniques of Artificial Intelligence (AI): Multiagent Systems and Reinforcement Learning. The system consists of four sub modules integrated.

The Information Collection Module is responsible for storing information generated by flight controllers. The Reinforcement Learning Module is responsible for system learning, which will receive information from collection module and transform into knowledge to be used in the future by specialist. The Forecast Scenarios Module is responsible for presenting the airspace scenario in a future instant T_{n+1} , in order to present to air traffic controller what might happen if he choose the action suggested by the system. The Decision Support Module will present the possible actions to be taken and every scenario that will be generated after take a certain action according with a prediction, including the impact on another airspace sector in the future. Figure 1 gives an overview of the architecture of the Decision Support Module and its interaction with the air traffic controller.

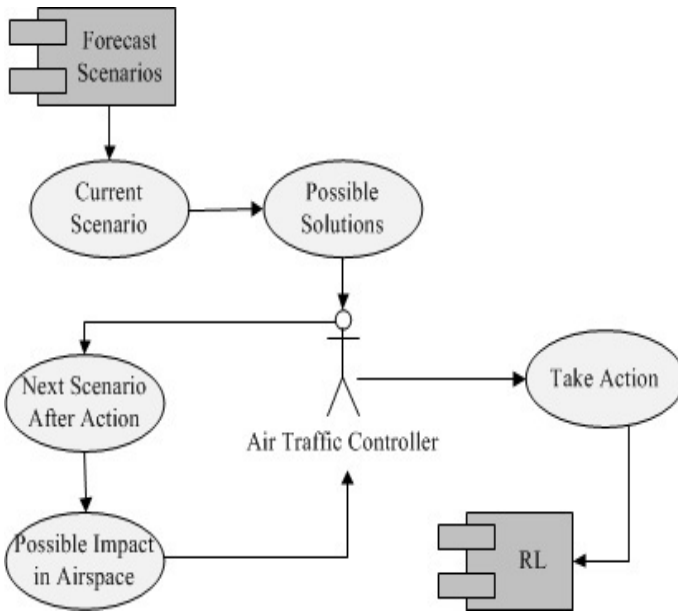


Fig. 1. AHPM architecture

This interaction between air traffic controller and AHPM can be understood as follows:

1. Current Scenario: It is responsible to display current scenario to air traffic controller.
2. Possible Solutions: It is responsible to display, based on Reinforcement Learning, possible scenarios considering taken actions in the past.
3. Next Scenario After Action: It is responsible for presenting the next scenario, if the chosen solution can be taken as restrictive measure.
4. Possible Impact in Airspace: It is responsible to evaluate and display possible impacts as congested or saturated sectors in airspace.
5. Take Action: It is responsible for receiving the action of the air traffic controller and send to the Reinforcement Learning Module for processing and storage. This information will be used for suggestions improvement in the future.

The Decision Support Module is the only module that will display the information and interacts with specialist. After presenting suggestions to air traffic controller the system will wait for his decision. If the chosen decision is accept the suggestions presented by the AHPM, these suggestions will be transformed into knowledge and stored in the database learning. If the specialist chooses an no listed action or only some of the suggested actions, this module will forward in the first case, to the Reinforcement Learning Module and in the second case,

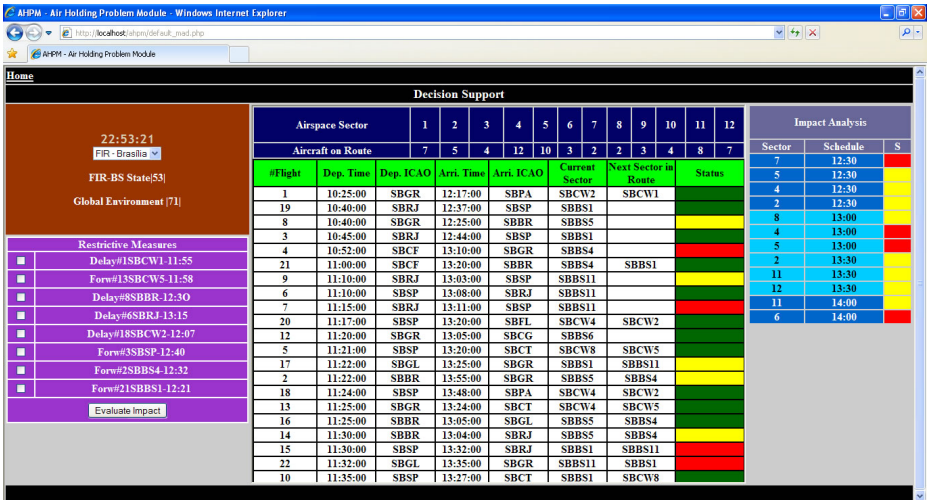


Fig. 2. Human-Computer Interaction with AHPM

for the Forecast Scenarios Module to recalculate states and forward to presenting the expected results for chosen actions to the air traffic controller.

The specialist will choose actions to be taken from possible solutions group. AHPM will try to predict the possible impact in airspace, if those actions are taken. Thus, the air traffic controller will decide about actions to be taken and the results will be storage in AHPM. The human-computer interaction is presented in Figure 2.

In the screen of AHPM is possible verify in a simple and clean manner how the air traffic controller will interact with the system. Initially, the system will verify the date and time which air traffic controller is running its activities. Thus, the screen will load automatically the flights that were planned for that time in an interval of ninety minutes and deviations that are occurring with a tolerance of three minutes. At this moment, the modules begin to act in a integrate way for presenting suggestions.

First of all, it is important to explain the left of screen. On top, it displays what is on FIR and analyzes its state. This is a state of Reinforcement Learning evaluation functions That Indicate an index level of air traffic Which is it defined the FIR state. The more near zero lower traffic congestion in the sectors of FIR on analysis. The global state follows the same principle but considers all sectors FIR's in airspace of Brazil.

On top middle, it displayed how many sectors and aircraft flying exist in analyzed FIR at this moment. There is a capability of aircraft for each airspace sector in one same moment. In Brazil, it is defined as congested sector if there are more than eleven aircraft in each sector and as saturated if more than thirteen aircraft. So, air traffic controller needs to analyze all this information in a short time and decide which are the best actions to airspace. It is presented twelve airspace sectors because this is the amount of sectors in FIR-BS.

On middle, it displayed all flights that are under responsible of a specific air traffic controller. It presents information such as flight number; departure time; ICAO code of departure; arrival time; ICAO code of arrival; current sector which aircraft is flying; if exists, next sector in route and air traffic status in the current airspace sector. The ICAO (International Civil Aviation Organization) code is an international identifier which is used for airports. In case of this status is green, the air traffic is fluent. If yellow, the sector probably will go turn congested and some action about this flight needs to be taken. If red, the sector probably will go turn saturated and needs some restrictive measures in flight with this status.

On bottom left, AHPM presents possible better restrictive measures to be taken at this moment over the flights under his responsibility. According to calculations made by Forecast Scenarios Module are identified some possible actions to be taken. These restrictive measures are classified into two types: delay and forward. One example of restrictive measures is '*Delay#4SBBR2-09:41*', which means delaying the entry of aircraft #4 in sector two of FIR-SBBR to 09:41. Another possible measure could be '*Forw9SBRJ-12:48*', which means forward the landing of aircraft #9 at the airport SBRJ to 12:48.

These restrictive measures are determined by Forecast Scenarios Module and take into consideration, basically all information presented on middle screen. These actions are suggestions for the air traffic controller, which can choose all, some or none. These suggestions consider several factors as system learning. The longer the system is in use, the best suggestions will be based on the scenarios like the current one.

After air traffic controller choose the actions to evaluate the impact, it will presented on top right the impact analysis. This analyze will show possible impacts in airspace sectors for the next three intervals of thirty minutes, for the sectors that will be affected in the FIR which is being analyzed. In case of this status is yellow, the sector probably will be almost congested in a determinate time. If red, the sector probably will be congested. It is possible to analyze the possible evolution by the three intervals, too. Thus, it is easier to verify if some action is so hard for a specific case.

The air traffic controller can analyses as many times as necessary, choosing different actions to be taken. When actions are taken in time, the results will be presented on central and it will be ready for air traffic controller starts the whole process again.

5 Conclusions

In the complex domain of ATFM, there are air traffic controllers who are responsible for some of the most critical activities, because requires a lot of concentration; air traffic experience; high commitment; ability to work under extreme pressure, among other factors that make the daily tiring and stressful. Besides the physical and psychological factors, there is the impact that their actions while working can cause in lives of so many people.

The artificial intelligence proves itself effective in helping the decision-making processes, specifically in the area of aviation. Due to factors such as acting in a

real time environment, using large amounts of data, lack of adequate experience to specialist, need to consider several factors in a short time, predict the impact before to take an action, and others. Systems that use one, or several, specific techniques of AI can address these needs and become in an important tool in ATM.

Although good solutions are built using the AI, it is required to be considered how will be held the interaction between system and expert. The DSS must evolve to the next level, which in addition to information for decision support the system must provide an efficient human-computer interaction. Currently, there are large amounts of data that can provide any information to the specialist, but there is so much information available that can limit the progress due to the difficulty of finding what it is important at the time required or by the poor interaction provided in the system.

The AHPM approach was proposed to support air traffic controller in decision-making process by the easy and fast interaction for all needed knowledge. Among some aspects proposed, it was possible to retain the knowledge of more experienced air traffic controllers in the system to help beginners; analyze and predict scenarios, within the time required to take a decision; assess potential impacts before taking a restrictive measure; and others actions to reduce holding traffic on the routes.

The AHPM gets to achieve a great level of human-computer interaction because the interaction is very simple and all the mandatory information to make great analysis is presented in the same screen. The information organization is clean and fast to find a specific data. This is especially important due to short time to detect problems, verify possible situations, analyze better actions to be taken and its possible risks.

For future work, we intend to perform the integration of the strengths of human-computer interaction of AHPM and currently used systems in Brazil to build a more efficient approach for air traffic controllers, such as the inclusion of radar maps with alerts messages to possible risk situations, the possibility of maps to present the 'Impact Analysis' and the inclusion of more information of other airspace control systems.

References

1. Agogino, A., Tumer, K.: *Learning Indirect Actions in Complex Domains: Action Suggestions for Air Traffic Control Advances in Complex Systems*. World Scientific Company (2009)
2. Arnowitz, J., Arent, M., Berger, N.: *Effective Prototyping for Software Makers*. Morgan Kaufmann Publishers (2007)
3. Bayen, A.M., Grieder, P., Meyer, G., Tomlin, C.J.: Lagrangian Delay Predictive Model for Sector-Based Air Traffic Flow. *AIAA Journal of Guidance, Control and Dynamics* 28(5), 1015–1026 (2005)
4. Beulens, A.J.M., Van Nunen, J.A.E.E.: The Use of Expert System Technology in DSS. *Decision Support Systems* 4, 421–431 (1988)
5. Bianco, L., Dell'Olmo, P., Odoni, A.R.: *New Concepts and Methods in Air Traffic Management*. Springer (2001)

6. Cherkassky, V., Mulier, F.: *Learning from Data: Concepts, Theory and Methods*, pp. 60–65. Wiley, New York (1998)
7. Cruciol, L.L.B.V., Weigang, L.: Air Holding Problem Solving by Reinforcement Learning to Reduce the Congestion in Airspace Sectors. In: *The 2012 International Conference on Artificial Intelligence, Proceedings of ICAI 2012*, pp. 272–278 (2012)
8. Dix, A., Finlay, J., Abowd, G.D., Beale, R.: *Human-Computer Interaction*, 2nd edn. Prentice-Hall (1998)
9. Grudin, J.: AI and HCI: Two fields divided by a common focus. *AI Magazine* 30(4), 48–57 (2009)
10. Lepreux, S., Abed, A., Kolski, C.: A human-centered methodology applied to decision support system design and evaluation in a railway network context. *Cognition Technology Work* 5, 248–271 (2003)
11. Leadbetter, D., Hussey, A., Lindsay, P., Neal, A., Humphreys, M.: Towards Model Based Prediction of Human Error Rates in Interactive Systems. In: *Australian Comp. Sci. Communications: Australasian User Interface Conf.* (2001)
12. Odoni, A.R.: The flow management problem in air traffic control. In: *Flow Control of Congested Networks*, pp. 269–288. Springer, Berlin (1987)
13. Palmer, E.M., Clausner, T.C., Kellman, P.J.: Enhancing Air Traffic Displays Via Perceptual Cues. *ACM Trans. Appl. Percept.* 5(1), 1–22 (2008)
14. Sutton, R.S., Barto, A.G.: *Reinforcement Learning: An Introduction*. The MIT Press, Cambridge (1998)
15. Wolf, S.R.: Supporting air traffic flow management with agents. In: *American Association for Artificial Intelligence Spring Symposium: Interaction Challenges for Intelligent Assistants* (2007)
16. Yoshikawa, H.: Modeling Humans in Human-Computer Interaction. In: Jacko, J.A., Sears, A. (eds.) *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*, pp. 118–149. Lawrence Erlbaum Associates, Mahwah (2003)

Decision Space Visualization: Lessons Learned and Design Principles

Jill L. Drury¹, Mark S. Pfaff², Gary L. Klein³, and Yikun Liu²

¹ The MITRE Corporation, Bedford, MA, USA
{jldrury,gklein}@mitre.org

² Indiana University Indianapolis, Indianapolis, IN, USA
{mpfaff,yikliu}@iupui.edu

³ The MITRE Corporation, McLean, VA, USA

Abstract. While the situation space consists of facts about what is currently happening, the decision space consists of analytical information that supports comparing the relative desirability of one decision option versus another. We have focused on new approaches to display decision space information that aids cognition and confidence. As a result of our earlier empirical work, we have developed a set of principles for visualizing decision space information. This paper describes those principles and illustrates their use.

Keywords: Decision space, situation space, option awareness, situation awareness, cognitive engineering, design principles, visualization.

1 Introduction

Professionals working in domains such as air defense, emergency response, and air traffic control must make decisions under complex situations. A core concern for all of these domains is the manner and extent to which decision makers develop option awareness (OA): the perception and comprehension of the relative desirability of the available options, as well as the underlying factors, trade-offs, and tipping-points that explain that desirability [27]. The number of options that decision makers need to consider, and the possible consequences of options that they need to weigh, can constitute a heavy cognitive burden. Thus there has been much interest in providing decision support systems (DSSs) that will act as cognitive prostheses to ease decision making.

Many DSSs provide displays that consolidate the facts of the situation. These displays are often called “dashboards” to emphasize the metaphor of the automobile dashboard, which shows speed, engine revolutions per minute, engine temperature, etc. The assumption behind a dashboard is that the relevant facts about a situation will lead to the best course of action [7]. Knowing the current state of the situation, which Hall, Hellar, and McNeese call the *situation space* [10], is undoubtedly necessary, but this information may or may not suggest options for handling the situation. Further, understanding the facts about the situation may not help a decision maker understand

the consequences of taking one option versus another, which Hall et al. call the *decision space* [10]. Understanding the situation space constitutes situation awareness (SA) [6], but understanding the decision space yields option awareness [27].

With sufficient OA, the decision maker is able to identify the most robust options: those which are most likely to turn out favorably under the widest range of possible future conditions [2, 20]. Robust options are in contrast to optimal options, which may appear desirable under specific predicted conditions, but can be sensitive to any discrepancies between the predictions and reality. We have been investigating generating OA through decision space visualizations (DSVs) that display trends, clusters of outcomes, goal conflicts, and tipping points between competing or synergizing options. DSVs are specific instances visual analytic displays and can be important components of DSSs.

We generate data for DSVs via exploratory modeling [2, 20], which uses computational simulations to generate all plausible outcomes of the available options under a broad range of input values and assumptions. Each iteration through the simulation model varies the assumed values of uncertain parameters through all combinations of plausible values that the decision maker cannot control. Taking a building fire as a simple example: What if the wind gets stronger? What if the storm arrives early? Visualizing the outcomes of thousands of rapidly computed experiments reveals whether the results of available options are sensitive to a particular underlying factor.

Our prior work shows that supporting OA using DSV improves decision accuracy and confidence over solely having situation space information [27]. The purpose of this paper is to formalize those results as a set of DSV design principles.

2 Related Work

There are many different sets of principles of human-computer interaction (HCI), such as Nielsen's ten usability heuristics [23], Norman's fundamental principles [24], and Shneiderman's "Eight Golden Rules" [30]. All of these sets of general HCI principles are intended to apply to almost any human-computer system. Examples of these principles are "be consistent," "provide visibility into system status," and "prevent errors." Our work does not aim to replace these general-purpose principles. Instead, we are *augmenting* general principles with those that are more specific to DSVs.

There is a long and diverse tradition of developing special-purpose principles or heuristics. Shneiderman's *Task by Data Type Taxonomy* includes seven high-level features highly salient to DSV: Overview, Zoom, Filter, Details-on-Demand, Relate, History, and Extract [31]. Norman developed principles for visual representations such as: match the properties of the visual representation with the information being represented [24]. Gerhardt-Powals' cognitive engineering principles are very relevant to visual analytics, and include: "reduce uncertainty," "fuse data," "group data consistently and meaningfully," and "automate unwanted workload." [9]. There is a lack of sets of principles, however, that specifically address visual analytic displays [13].

While many DSSs solely provide situation-space information, some do provide explicit support for exploring options. One example is RODOS, a DSS used for nuclear

remediation management [8]. In its original form, it provided point estimates through stacked bar charts, indicating the probable outcome of each option and the relative contribution of its attributes. Preliminary support was added for visualizing uncertainty [28], but only by adding stacked bars for the 5th and 95th percentile outcomes (suggesting the best and worst cases) on either side of each of the deterministically calculated outcomes. Unfortunately, this approach still obscures important complexity hidden in the distribution of modeled outcomes, such as skewness, clustering, and multimodality. Embedded in these outcome distributions is vital information about the interacting influences of the complex and uncertain underlying factors influencing the outcome of a given option.

3 Developing the Principles

The following design principles for DSV represent a synthesis of empirical research on DSSs, information visualization, and interaction design, supported by our research including two computational [19, 22] and five human-subjects experiments [27, 21]. Complete details of the methods and results of these experiments have been previously published in the citations provided, so will not be repeated here.

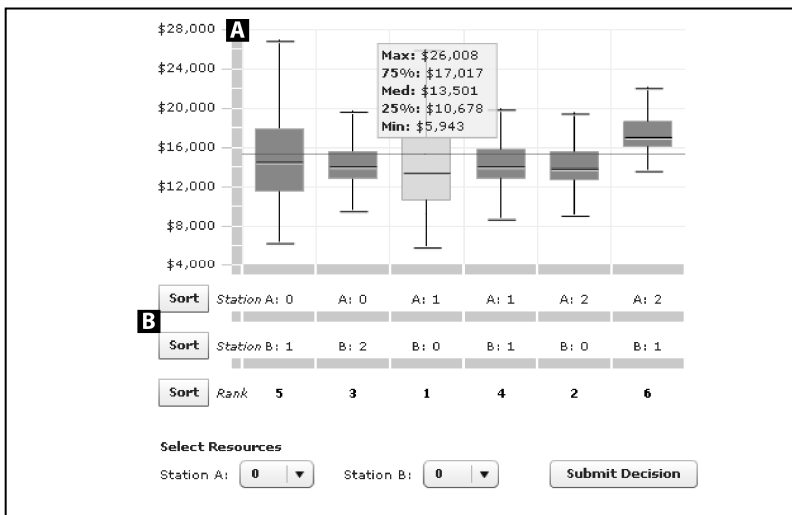


Fig. 1. User interface example showing (A) frequency distributions of outcomes for top six options, and (B) sorting by rank or quantity of resources. Note that boxplots summarize the distribution of outcomes, with one boxplot for each option.

For the human-subjects experiments, we designed a series of interfaces to provide an experimental problem space with a plausible amount of urgency and uncertainty using relatively familiar scenarios (robberies, accidents, fires, etc.) and resources (police cars, ambulances, fire trucks). We used the NeoCITIES emergency response simulation model [11] to predict the range of outcomes for the available options.

Figure 1 shows a portion of the DSV used in two experiments [26]. The primary components are noted with capital letters. The available options (in this case, various combinations of fire trucks from two different stations) are along the horizontal axis, and the evaluative metric (in this case, combined cost of material damage, casualties, and expended resources) is along the vertical axis (A). Users could sort options by quantity or rank, with rank 1 being the most robust option (B).

The principles were developed during the course of our experiments. The results of these experiments progressively refined the fundamental HCI principles listed above, resulting in the seven principles of DSV presented in text boxes below.

4 Principles of Decision Space Visualization

1. Allow users to apply their own mental models to their situational observations and provide input parameter values. Do *not* require users to set input parameter values for information that can be accurately and automatically obtained elsewhere.

In our second experiment [5], we introduced controls for users to set input parameter values for the predictive model based on their interpretation of the magnitude and impact of the emergency scenario. In that experiment, even though the user-provided values did not actually influence the model, the act of providing the inputs resulted in a significant improvement in decision confidence when using the DSV compared to our first experiment [4]. These results are congruent with those of Shneiderman [32], which showed that users who actively interact with data have more confidence than those who do not have as much interaction.

Of course, users should not be asked to enter data that can be acquired automatically, such as current weather, traffic, or other information that could be retrieved from live streams or databases. Having the system populate the relevant fields is an example of appropriately applying Nielsen's heuristic for efficient interface use [23].

2. Allow the user to apply their real-world knowledge to set weights or values for the scoring function (the criteria for ranking the options).

One of the main interface additions of interest in our fourth experiment [26] was a set of customizable controls for users to manipulate the parameters of the scoring function. Changing the weighting strategies could alter the rankings of the options by varying the amount of weight given to each of the five parameters of the boxplots. Compared to the preceding experiment without these weighting controls and using the same set of emergency scenarios [25], participants made significantly more accurate decisions with higher confidence.

3. Provide an overview of the top several options and allow users to employ their pattern recognition, judgment, and values to choose the desired option.
 - a. Do not have the visualization identify a single, firm recommendation to users.
 - b. Do not provide a very large number of options, especially when many of them are much less desirable and thus unlikely to warrant serious consideration.

One of the key properties of visual analytics tools is an overview to help the user develop overall awareness of the information presented [15]. These techniques help decision makers to identify anomalies, trends, and patterns by leveraging the strengths of their visual perception for perceptually tractable parallel comparison of options. Therefore, we employed this principle in all of our visualizations.

Many DSSs provide a recommendation for a single option that is optimal if all of the assumptions underlying the recommendation hold true. The difficulty with this approach in complex, uncertain, real-world situations is the impossibility of predicting the future with 100% certainty. If the assumptions prove invalid, or uncontrollable conditions emerge, then the “optimal” solution may perform quite poorly. Experienced decision makers often distrust decision support systems after seeing recommendations that did not perform well [12]. Even when DSSs work well, they are often not used because decision makers do not know how the systems arrive at their recommendations [17]. Our DSV provided multiple options that are explored under all plausible assumptions so that decision makers could use their own judgment.

Our interface only displayed the six currently top-ranked options based on well-known limitations of short-term memory (the “magical number 7” [1]). Displaying too many options, especially those that perform poorly and so are unlikely to be chosen, can simply overwhelm users (e.g., see [29]).

4. When constructing DSVs, trade off unnecessary fidelity in favor of speed of response. Determine needed fidelity level based on whether DSVs generated from models of a lower fidelity level would lead to the same decision as DSVs constructed from data obtained via a higher-fidelity model.

Even with modern computer processors, the models that are instrumental in robust decision-making (RDM) analyses and that produce the data required by DSVs can be computationally challenging. We envision real-time or near-real-time support to users, which requires response times that are acceptable to them. When we were faced with a model that took days to run, we performed a computational study to determine the level of fidelity and precision that was truly needed [22].

Accordingly, we compared the top-ranked options of several different pandemic influenza response models. We compared 16 courses of action from four different influenza models (varying in precision and fidelity) and discovered that the models generally agreed on the top-ranked options, but dramatically disagreed on the lowest-ranked options [22]. This was a valuable finding in practical terms, since the user will

examine the best options available, not the worst ones. Not only did this finding confirm that presenting additional options provided diminishing returns, but revealed that using the RDM methodology allowed faster-running, lower-fidelity models to provide essentially the same recommendations as more computationally-intensive, high-fidelity models.

5. Show the consequences of choosing one option versus another under a variety of possible conditions rather than a single set of “most likely” conditions.
 - a. Use a frequency-based presentation, not a probability-based presentation.
 - b. Reveal the shapes of the distribution of outcomes.

While Principle 3 is concerned with showing multiple options to enable visual comparison of outcomes *between* options, Principle 5 is based on showing multiple outcomes *within* each option. Such visualizations take advantage of humans’ innate perceptual abilities to find patterns, with the advantage that simulating many “what if?” cases has been offloaded to the computer instead of requiring mental simulation.

By showing the range of possible results from choosing a particular option, the DSV takes into account the uncertainty regarding conditions outside of decision makers’ control. How this uncertainty data is presented can make a difference in users’ level of understanding, however [16]. Only showing averages or means conceals critical details about the complexity and uncertainty underlying each of the available options. To compare the robustness of several options, each with a distribution of possible outcomes, outcomes can be mapped onto a single user-selected multi-attributed cost metric. To combat biases known to result from displaying point estimates of probabilities [16], our DSV designs use a frequency format to display the results.

6. Provide interactive filtering and sorting for viewing subsets of the data that underlie the decision space.

Others have demonstrated how direct manipulation of the interface helped users rapidly perform and refine dynamic queries of a database, with participants successfully finding trends and anomalous data [33]. Direct manipulation of a dynamic interface refers to immediately visible user-driven adjustments to search results or other data using interactive controls, such as sliders or buttons. Applied to searching the decision space, users choosing an option will often need to dynamically filter, sort, and drill down deeper into the exploratory modeling results.

7. Support comprehension of the factors and relationships mediating the consequences of choosing one option versus another.

We have termed this aspect of the DSV *option awareness level-2*, or OA2 [18]. Supporting this aspect of DSV requires effective multi-factor comparisons for

uncovering the reasons for the relative robustness of competing options. In our latest interface [3], users can adjust a threshold to differentiate between favorable and unfavorable outcomes on the vertical cost axis of the DSV overview, which now shows outcomes as individual points in addition to box plots. As they adjust this threshold, the system dynamically renders a tree of the causes and conditions that explains the differences between favorable and unfavorable outcomes.

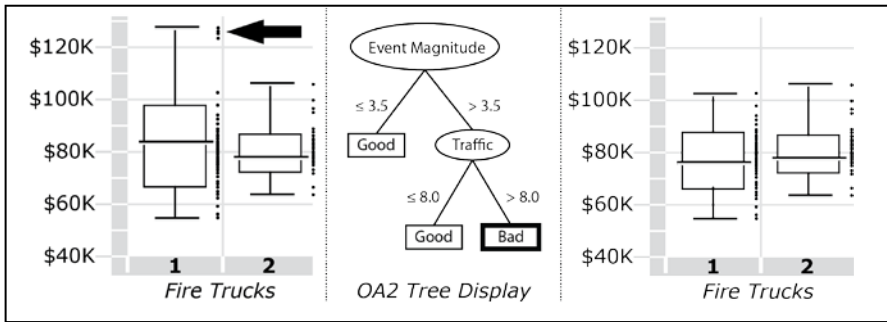


Fig. 2. Simplified example, from left to right, of a user identifying an anomalous cluster of especially poor outcomes, drilling down into the underlying causes explaining that cluster, and addressing the traffic problem (the cause of the poor outcomes), resulting in a revised decision space.

5 Using the Principles

Figure 2 shows an abstraction of a fire dispatcher developing OA with DSV support. She begins by entering the reported magnitude and location of the fire (Principle 1). The system automatically determines the weather conditions at the fire location, since winds and precipitation can affect the fire's behavior.

The decision space overview is on the left in Figure 2 (Principle 3). This overview appeared quickly because the model driving the DSV is at the lowest reasonable level of fidelity (Principle 4). In the overview, the two box-plots provide a frequency distribution (Principle 5a), with the shapes of the distributions depicted using scatterplots of dots showing the approximate numbers of potential cases at each cost point (Principle 5b). The overview shows that sending one truck has the possibility of being less expensive than sending two (because it costs more to send two trucks than one) but that sending one truck also has the possibility of being *more* expensive than sending two (because one truck may not be sufficient to put out the fire). In other words, sending two trucks has a worse best-case cost, but a better worst-case cost (Principle 5). While only two options are being shown in this brief illustration, the fire dispatcher is viewing six options (Principles 3a and 3b), and explores the options by sorting them so that the option with the lowest best-case cost is shown on the far left (Principle 6).

The arrow on the left hand side of Figure 2 points towards an anomalous cluster of bad (high cost) outcomes, making option 1 look undesirable. The user selects this cluster for further inspection (Principle 7). The tree in the middle of Figure 2 shows

the hierarchy of factors leading to the selected “Bad” outcomes: these occur when the event magnitude is greater than 3.5, and traffic is greater than 8.0 (the units are arbitrary in this example). Knowing that traffic is a critical factor leading to that cluster of bad outcomes, the user generates a new option by adding a police car to control traffic, producing the updated decision space on the right. Option 1 is now more robust than option 2 because it reduces the cost to the city and preserves limited resources for future events. Pre-populating the model with such opportunities for synergy, drawn from experienced decision makers, will shorten the time to search the decision space for the most robust solutions in scenarios with far more competing factors than this simple example.

5 Discussion

In summary, the first two principles pertain to the collaborative partnership between the human decision maker and the computational modeling of the decision space. Principles 3 – 5 describe critical considerations for user-centered visualization of the decision space. The last two principles suggest methods to support users developing deeper option awareness by interacting with and drilling down into the decision space.

Making underlying factors explicit and explorable in the DSV has significant benefits at the individual and team level. This knowledge helps individuals develop accurate mental models of the decision space. The accuracy of mental models is vital, since decision makers do not solve the actual problem, but solve their mental model of the problem; ideally, the two are similar enough to produce successful outcomes. At the team level, this knowledge can point out opportunities for collaboration, as well as goal conflicts within or between teams. In addition, to the extent that using the DSV provides distributed teammates with a similar mental model of the decision space, that shared mental model is vital for effective communication and team effectiveness [14].

Making the computer part of the team is embodied in the principles that we have laid out. We hope these principles for designing decision spaces will eventually lead to more robust decision making across a wide array of applications.

Acknowledgments. We thank Aeshvarya Verma and Sung-Pil Moon from Indiana University Indianapolis, and Loretta More from Pennsylvania State University. We also thank the participants in the experiments. This work was supported in part by MITRE Corporation Innovation Projects 43MSR001-EA, 45MSR026-FA, and 43MSR003-KA.

References

1. Baddeley, A.: The magical number seven: Still magic after all these years? *Psychological Review* 101(2), 353–356 (1994)
2. Chandrasekaran, B.: From optimal to robust COAs: Challenges in providing integrated decision support for simulation-based COA planning, Laboratory for AI Research. The Ohio State University (2005)

3. Drury, J.L., Klein, G.L., Musman, S., et al.: Requirements for data mining the decision space. In: Proc. of the 2012 International Command and Control Research and Technology Symposium (ICCRTS), Fairfax, VA (2012)
4. Drury, J.L., Klein, G.L., Pfaff, M.S., et al.: Data visualizations for dynamic decision support. In: Human Interaction with Intelligent and Networked Systems Workshop, Proc. of the Intelligent User Interfaces Conference (IUI 2009), Sanibel Island, FL (2009)
5. Drury, J.L., Klein, G.L., Pfaff, M.S., et al.: Dynamic decision support for emergency responders. In: Proc. of the 2009 IEEE Conference on Technologies for Homeland Security (HST 2009), May 11-12, pp. 537-544 (2009)
6. Endsley, M.R.: Toward a theory of situation awareness in dynamic systems. *Human Factors* 37(1), 32-64 (1995)
7. Few, S.: *Information dashboard design: The effective visual communication of data*. O'Reilly Media, Inc., Sebastopol (2006)
8. Geldermann, J., Bertsch, V., Treitz, M., et al.: Multi-criteria decision support and evaluation of strategies for nuclear remediation management. *Omega* 37(1), 238-251 (2009)
9. Gerhardt-Powals, J.: Cognitive engineering principles for enhancing human - computer performance. *International Journal of Human-Computer Interaction* 8(2), 189-211 (1996)
10. Hall, D.L., Hellar, B., McNeese, M.D.: Rethinking the data overload problem: Closing the gap between situation assessment and decision making. In: Proc. of the 2007 Symposium on Sensor and Data Fusion (NSSDF) Military Sensing Symposia (MSS), McLean, VA (2007)
11. Hamilton, K., Mancuso, V., Minotra, D., et al.: Using the Neocities 3.1 Simulation to Study and Measure Team Cognition. In: Proc. of the Human Factors and Ergonomics Society Annual Meeting, vol. 54(4), pp. 433-437 (2010)
12. Hammond, K.R., Summers, D.A., Deane, D.H.: Negative effects of outcome-feedback in multiple-cue probability learning. *Organizational Behavior and Human Performance* 9(1), 30-34 (1973)
13. Hanrahan, P., Eick, S., Ebert, D.S., et al.: Visual representations and interaction technologies. *Illuminating the Path: The Research and Development Agenda for Visual Analytics* (2005)
14. Hinsz, V.B.: Metacognition and mental models in groups: An illustration with metamemory of group recognition memory. In: Salas, E., Fiore, S.M. (eds.) *Team Cognition: Understanding the Factors that Drive Process and Performance*, pp. 33-58. American Psychological Association, Washington, DC (2004)
15. Hornbæk, K., Hertzum, M.: The notion of overview in information visualization. *International Journal of Human-Computer Studies* 69(7&8), 509-525 (2011)
16. Ibrekk, H., Morgan, M.G.: Graphical communication of uncertain quantities to nontechnical people. *Risk Analysis* 7(4), 519-529 (1987)
17. Kayande, U., De Bruyn, A., Lilien, G.L., et al.: How incorporating feedback mechanisms in a DSS affects DSS evaluations. *Information Systems Research* 20(4), 527-546 (2009)
18. Klein, G.L., Drury, J.L., Pfaff, M.S., et al.: COAction: Enabling collaborative option awareness. In: Proc. of the 15th International Command and Control Research and Technology Symposium (ICCRTS), Santa Monica, CA (2010)
19. Klein, G.L., Pfaff, M.S., Drury, J.L.: Supporting a robust decision space. In: Proc. of the AAAI Spring Symposium on Technosocial Predictive Analytics (AAAI-TPA 2009), Palo Alto, CA (2009)
20. Lempert, R.J., Popper, S.W., Bankes, S.C.: *Shaping the next one hundred years: New methods for quantitative, long-term policy analysis*. RAND, Santa Monica (2003)

21. Liu, Y., Moon, S.P., Pfaff, M.S., et al.: Collaborative option awareness for emergency response decision making. In: Proc. of the 8th International Conference on Information Systems for Crisis Response and Management (ISCRAM), Lisbon, Portugal (2011)
22. Mathieu, J., Pfaff, M.S., Drury, J.L., et al.: Tactical robust decision making methodology: Effect of disease spread model fidelity on option awareness. In: Proc. of the 7th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2010), Seattle, WA (2010)
23. Nielsen, J.: Heuristic evaluation. In: Nielsen, J., Mack, R.L. (eds.) Usability Inspection Methods, pp. 25–62. John Wiley & Sons, Inc., New York (1994)
24. Norman, D.A.: The design of everyday things. Basic Books, New York (1988)
25. Pfaff, M.S., Drury, J.L., Klein, G.L., et al.: Decision support for option awareness in complex emergency scenarios. In: Proc. of the 3rd International Conference on Applied Human Factors and Ergonomics (AHFE), Miami, FL (2010)
26. Pfaff, M.S., Drury, J.L., Klein, G.L., et al.: Weighing decisions: Aiding emergency response decision making via option awareness. In: Proc. of the 2010 IEEE International Conference on Technologies for Homeland Security (HST), November 8–10, pp. 251–257 (2010)
27. Pfaff, M.S., Klein, G.L., Drury, J.L., et al.: Supporting complex decision making through option awareness. *Journal of Cognitive Engineering and Decision Making*, Advance Online Publication (2012)
28. Raskob, W., Gering, F., Bertsch, V.: Approaches to visualisation of uncertainties to decision makers in an operational decision support system. In: Proc. of the 6th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2009), Göteborg, Sweden (2009)
29. Sethi-Iyengar, S., Huberman, G., Jiang, W.: How much choice is too much? Contributions to 401(k) retirement plans. In: Mitchell, O.S., Utkus, S. (eds.) *Pension Design and Structure: New Lessons from Behavioral Finance*, pp. 83–95. Oxford University Press, Oxford (2004)
30. Shneiderman, B.: *Designing the user interface: Strategies for effective human-computer interaction*, 3rd edn. Addison-Wesley, Reading (1993)
31. Shneiderman, B.: The eyes have it: A task by data type taxonomy for information visualizations. In: Proc. of the IEEE Symposium on Visual Languages, pp. 336–343 (1996)
32. Shneiderman, B.: Direct manipulation for comprehensible, predictable and controllable user interfaces. In: Proc. of the 2nd International Conference on Intelligent User Interfaces, Orlando, Florida, United States, pp. 33–39 (1997)
33. Williamson, C., Shneiderman, B.: The dynamic HomeFinder: Evaluating dynamic queries in a real-estate information exploration system. In: Proc. of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 338–346 (1992)

The Language of Motion: A Taxonomy for Interface

Elaine Froehlich, Brian Lucid, and Heather Shaw

efroehlich@cox.net,
brian.lucid@gmail.com, heathershawdesign@gmail.com

Abstract. This project presents a taxonomic tool for designing with motion. Paul Klee dedicated his life to the study and teaching of motion. "I should like to create an order from feeling and, going still further, from motion." [1] The natural state of interaction with digitized information includes motion. Our human brains have evolved physiological systems and organic structures that respond instinctively, tuned to organic motion. This human bias toward organic, natural motion presents opportunities for the use of motion in interfaces. Using motion in computing devices inspired by the natural world will create deeper, more emotionally engaging experiences. This study focuses on understanding the basic elements of motion in order to use it as a component in the design of digital interfaces. It presents a taxonomy of motion with the goal of describing fundamental qualities of motion used in the 2-dimensional, framed space of a screen: screen position, direction, principles, attributes and the resulting behaviors that can be created using them. The documentation presented defines a language for motion in interface. The taxonomy was built on discrete gestural motion videos taken from nature. The video segments are limited to short motions that show a complete but definable idea. The videos tend to be a few seconds in length though a few of them take several seconds to complete their motion idea.

Keywords: Dynamic media, motion design, motion, interface, screen area, direction, principles, attributes, behavior, taxonomy.

1 Why Understand Motion?

This project emerged from a desire to use motion as an element in my own interface design practice. Motion in this case meaning motion within the framed space of the screen. The screen does not move, as a camera might move with the action in a film. Certain kinds of motions have long been used in interfaces to indicate functions taking place: progress bars to indicate a process is under way, for example.

A while ago, a web-based product development application that I was designing, calling for information displayed in three levels of detail, posed a problem difficult to address with static tools. The display required areas showing complete detail, partial detail and abbreviated detail, a level that displayed no more than an indication of change in a database. Most data sources in this application used a numerical chart that changed by incrementing numbers to indicate changes in the abbreviated level data. If the data source had many changes, the motion of the incrementing numbers made the

change visible. During periods when few updates were entered in the database, the numerical value change was the only indication that activity was taking place in that data source. As an information display, it was easily overlooked. Motion of some sort seemed a natural fit to express the idea of change taking place in the database; difficult to address using the tools available at that time.

1.1 Motion Exists in Time

Time permits the perception of motion. In “From Eternity to Here,” [2] Sean Carroll defines time as an increment, a definable point, and a medium through which we move.” On planet earth we count time as a fragment of each daily revolution. Visualizing the passing of increments can show time as we conceive of it.

The human perception of time is influenced by psychological perspective more than rational perception. Another approach for visualizing time is by interval. Rather than counting durations into smaller slices, relationships of intervals compared against a mean or against other intervals shows time through intersection or separation of motions. By considering intervals in relation to each other, a different kind of information about time may be communicated that expresses the human perception of time in the communication experience of the interface.

2 The Essential Four Components

Four components define the essential definitions within the taxonomy. Screen position and the direction of motion comprise the obvious areas to start defining screen motion. Principles and Attributes refine and enrich motion messages.

2.1 Screen Position

Screen position refers to the area or areas of the screen where motion occurs. As with any interface, distribution of screen elements allows content and control areas to be easily distinguished.

Screen. The placement of motion as meaningful elements would take advantage of screen area to refine the meaning of those motions. The screen will contain areas where motion is located and in most cases, still areas. Moving spaces on the screen can be categorized. Motion may be localized to an area, may cover the whole screen, several areas or it may move from quadrant to quadrant. The motion may occur along the edges or take place in the center.

Whole screen

Part of screen

Center of screen

Edges of screen



Fig. 1. Whole screen



Fig. 2. Part of screen (bottom edge)

Screens of today cover a wide range of sizes and styles. Preference here was given to breaking the screen into thirds, allowing the definition of multiple areas within the videos. Limiting the number of areas maintains simplicity when analyzing the motion videos. This taxonomy does not differentiate for the screen size. In practice, a larger screen might be broken down into more areas. This analysis looks at the screen cut into thirds horizontally, vertically, in dimension on the z-axis and at geometric shapes that motion can take on the screen.

Position

- Left, center, right
- Top, middle, bottom
- Foreground, middle ground, background



Fig. 3. Position: left, center, right

Geometric shapes

- Oblique Square Triangle
- Circle Spiral Radial



Fig. 4. Shape: circle

2.2 Direction

Motion implies direction. Regardless of the amount of space motion takes up on screen, or its speed, it will have a direction within the frame. Direction pertains to the orientation of the motion relative to the screen: up, down, left, right, toward, away from, at diagonals, concentric (toward center), eccentric (away from center); straight or turn; variables within direction; and combinations of the above.



Fig. 5. Direction: Right

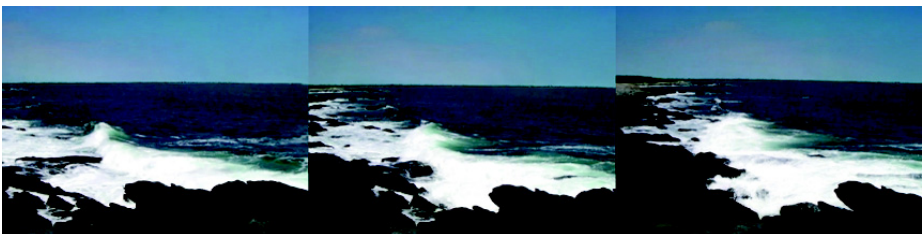


Fig. 6. Direction: Left



Fig. 7. Direction: Away

2.3 Principles

In design as visual language, we use principles to allow us to isolate ways of identifying visual components into definable abstract ideas. Principles reflect back to the basic design theory: rhythm, texture, pattern, contrast, repetition, that may be stationary or moving, and sequence, interval, velocity, synchronization, pace, transition, etc. that require change over time to reveal themselves.

Identifiable sequences described as principles can be applied to still or mobile examples equally well. When the element of time is part of the example these principles become the building blocks for motion ideas used to create visual communication.



Fig. 8. Principle: contrast of direction

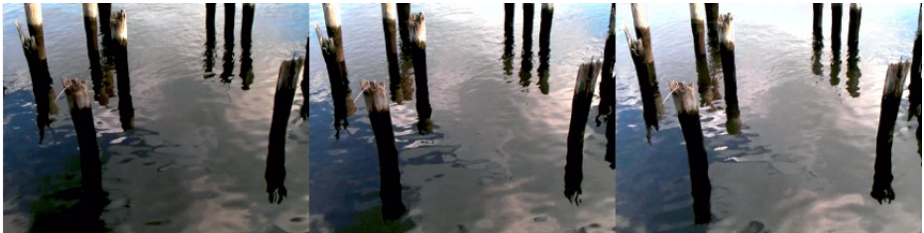


Fig. 9. Principle: texture

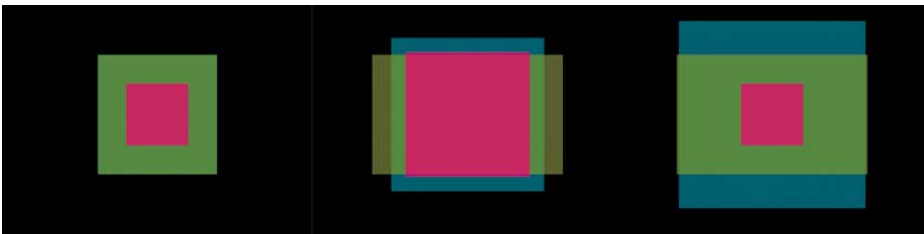


Fig. 10. Principle: interval

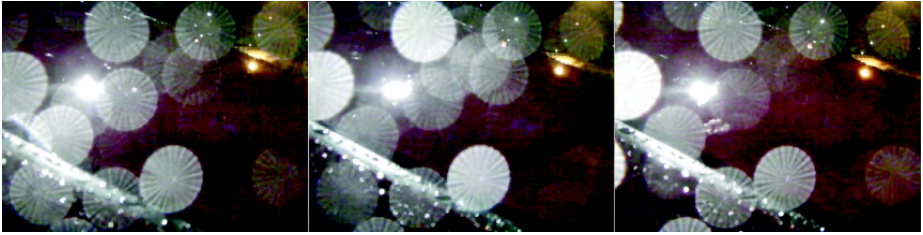


Fig. 11. Principle: pattern



Fig. 12. Principle: synchronous



Fig. 13. Principle: asynchronous

2.4 Attributes

Attributes to address the quality of the motion as it appears. Attributes exist in oppositional pairs. The quality of motion on the screen manifests in multiple ways. Understanding the attributes benefits when they are evaluated in contrast with opposite attributes.

Oppositional pairs:

- Proximity: together, apart
- Density: consolidated, dispersed
- Depth: pass in front, behind
- Distance: near, far away
- Quantity: single object, multiple objects



Fig. 14. Attribute: single object, multiple objects

Scale: toward, away from
Coincidence: before/after, during/simultaneously
Size: large elements, small elements move
Speed: fast, slow

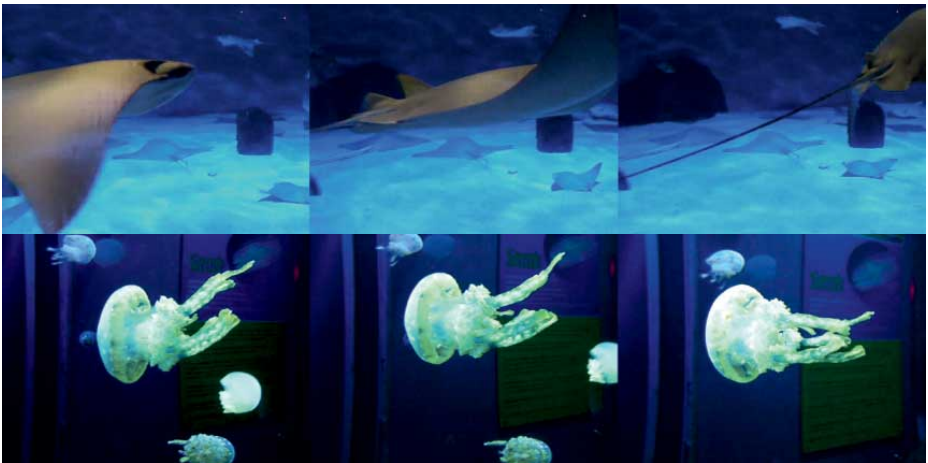


Fig. 15. Attribute: fast, slow

Noticeability: obvious, subtle
Change: change of focus, direction, quantity, ...
Causality: cause, effect

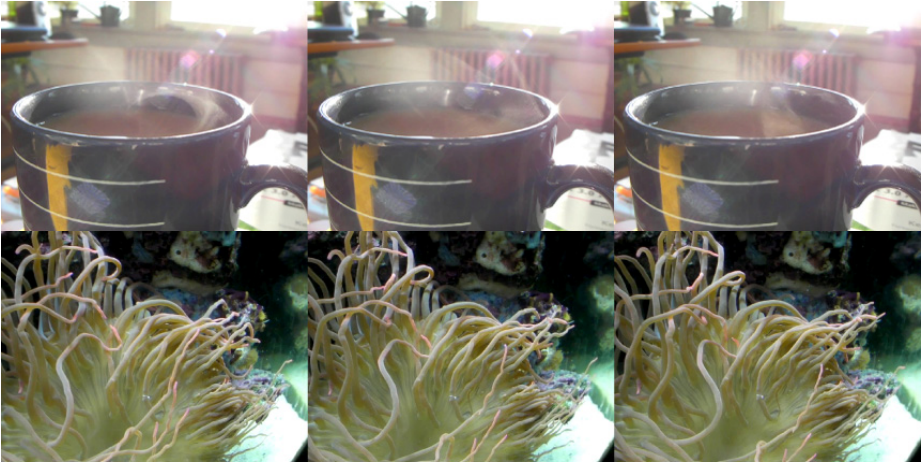


Fig. 16. Attribute: obvious, subtle



Fig. 17. Attribute: cause and effect

3 Behavior

In the world, behaviors communicate intention and physical non-intention, or the logical conclusion of a gestural arc. Motion with meaning on screen can be defined as behavior. Something that goes from top of the screen to stop at the bottom is falling. The characteristics of its landing tell much about the quality of that fall. If the same thing travels from the top to the bottom of the screen and springs back up to the top, it would be bouncing. How those behaviors become articulated in an information system results in the user's understanding. Fall and bounce carry different connotations.

To create behaviors, combining elements from the essential four parts of the taxonomy builds behaviors. Richer dimension to the types of motions created with the understanding of these categorical definitions allow interface designers to mix motions without relying on narrative to create meaning within their products.



Fig. 18. Behavior: wave



Fig. 19. Behavior: bob



Fig. 20. Behavior: turn

4 Motion or Experience Design

The earliest Graphical User Interfaces included motion as part of their visual display. Though they were limited by the processing power of those early machines, interfaces used motions to convey certain kinds of messages: show, tell, orient, acquaint or warn.

Interface motion currently follows a few well-saturated forms. One common current use of motion maintains the orientation of the user as the interface transitions between states. Grafting motion into existing static interfaces is doomed to failure. Imagine the usability problems of a Microsoft Word with a motion-based interface.

Motion design informed by understanding the motion of natural environments has a place in interface design. Motion presents a rich area for solving interface challenges. As robust data collection allows flowing data analysis, motion carefully articulated for meaning could be used to find patterns within the flow otherwise invisible to static visualization. As more and more products incorporate user interfaces, a unique

motion scheme defining the personality of the product could differentiate its brand.

Newer devices, multi-user spaces or ambient components of complex systems will benefit from the use of motion. Those emerging environments will demand solutions to problems that traditional interfaces never faced. The inclusion of motions created through interacting with them may be required when more than one user is interacting on a single screen.

This work touches the surface of a broad and deep topic. The presentation of the taxonomy creates an environment for understanding motion on the screen. This taxonomy is not exhaustive. Exploration of more motion from nature, investigation into the use of moving textures and patterns as meaningful elements of moving interfaces promises exciting opportunities to leverage into new types of interaction.



Fig. 21. Principle: rhythm (light)

References

1. Klee, P.: *The Thinking Eye* by Jurg Spiller, 2nd revised edn. Percy Lund Humphries & Co. Ltd. (1964)
2. Carroll, S.: *From Eternity to Here: the quest for the ultimate theory of time*. Plume, New York (2010)

Adaptive Consoles for Supervisory Control of Multiple Unmanned Aerial Vehicles

Christian Fuchs¹, Sérgio Ferreira², João Sousa³, and Gil Gonçalves²

¹ Faculty of Aerospace Engineering, Delft University of Technology

² Department of Informatics Engineering,

³ Department of Electrical and Computer Engineering

School of Engineering, Porto University (FEUP)

c.fuchs@student.tudelft.nl, {asbf,jtasso,gil}@fe.up.pt

Abstract. With the prevailing increase of complex operational scenarios, involving multiple unmanned aerial vehicles (UAV), the concerns with the natural increase of operator workload and reduction of situational awareness have become paramount in order to safeguard operational security and objective completion. These challenges can be tackled through alterations of the autonomy levels of the vehicles, however this paper explores how these issues can also be mitigated by changing the way information is presented to the human operator. Relying upon an established framework, that supports operational scenarios with multiple UAVs, a series of display alterations were performed to existing operation consoles. After test sessions, in a simulated environment, with human participants of different levels of operational certification, feedback and results are distilled and analysed. Operator feedback demonstrated an overwhelming preference for the developed consoles and results showed an improvement of situation awareness, as well as reduction of workload.

Keywords: Operator, Situational Awareness, UAS, UAV, Workload, Command and Control, Interface.

1 Introduction

Recent years have witnessed unprecedented technological developments in computing, communications, navigation, control, composite materials and power systems. These developments have allowed the design and deployment of a multitude of extremely capable unmanned aerial vehicles (UAV) and unmanned aerial systems (UAS). As the operational capacity of UAS continues to grow, these systems can include multiple UAVs operating as a team, furthermore solidifying their employment in military and civilian scenarios. This causes an increase of the workload felt by the human element of these UASs, as well as a decrease in their situational awareness during the operation.

Normally workload and awareness issues are handled by changing the vehicles autonomy levels, increasing them in order to ease the human operator's experience. However we propose that changes made to the information's layout, and to the manner in which it is conveyed to the human operator, provide us a tool with which to affect operator workload and awareness in a positive fashion.

2 Method

This work was conducted at the Underwater Systems and Technology Laboratory (LSTS) as part of the work developed through the PITVANT project. At the LSTS we have been designing, building and operating a significant number of heterogeneous unmanned vehicles. These include Remotely Operated Vehicles (ROV), Autonomous Underwater Vehicles (AUV), Autonomous Surface Vehicles (ASV), and UAVs as a result of our collaboration with the Portuguese Air Force Academy. Furthermore we made extensive use of the LSTS's existing toolchain [1] for control and development comprised by the C4I (Command, Control, Communications, Computer and Intelligence) system Neptus [2], the vehicle task manager, control and navigation software DUNE (Dune Uniform Navigational Environment) and the IMC (Inter-Module Communication) communication protocol [3]. Since it is already amply used by both the Portuguese Navy and Portuguese Air Force Academy the toolchain allows us to receive a great amount of feedback and gives us access to a large number of potential test subjects.

2.1 Console Profiles

In order to adapt the console to the specific requirements of a situation, the concept of console profiles is introduced: A console profile is a predefined set of display elements which is geared towards a specific task. It is then possible to switch between profiles during a mission, either manually or automatically.

2.2 Operator Survey

In the beginning, several certified UAV operators are surveyed. They are asked what information an operator does or does not need to see, how much control he desires to have over the UAV, in different scenarios, and where his focus lies. Each of those questions is answered for 4 different tasks:

- Controlling a single UAV;
- Controlling multiple UAVs;
- Operating an onboard video camera;
- Operating as a tactical commander.

Based on this information, a decision is made regarding what elements to include or omit in each console profile.

2.3 Test Setup

As a first step, workload and situational awareness are evaluated in a simulated environment. During this test, the operator is asked to control an increasing number of UAVs and execute tasks such as changing flight plans, airspeeds and altitudes. The location and tasks to be executed are equal to those encountered in numerous previous flight tests performed at Ota airfield, Portugal.

Table 1. Questions asked during the test to assess operator situational awareness. Questions 12 was not asked as part of SAGAT but noted without the participants' knowledge.

# Question	
1 How many UAVs are you controlling?	7 What is the heading of each UAV?
2 Which UAVs are those?	8 What are the UAVs' position relative to each other?
3 What is the main UAV?	9 What part of the plan are the UAVs executing now/next?
4 What is the altitude of each of the UAVs?	10 What is the status of each UAV?
5 What is the airspeed of the main UAV?	11 What were you last orders?
6 Where on the screen are the UAVs?	12 How many anomalies were detected?

To compensate the lack of naturally occurring stress in an operational scenario, inherent to having real hardware that would be lost in case of a catastrophic failure, the number of UAVs to be controlled, as well as the number and frequency of ordered tasks, are increased significantly.

Even though 4 different profiles were created, this test concentrates on the control of a single UAV and multiple UAVs, therefore only the profiles for single and multi UAV control are used.

Two different measurement techniques are used to judge the operator's workload and situational awareness: NASA TLX [4] and SAGAT [5], respectively. Additionally, the participants are asked to point out any anomalies they encounter. These include a sudden change in altitude/airspeed or subsystem failures. A summary of the questions is given in Table 1, while Table 2 shows when each measurement was taken.

Table 2. Test scenario showing how the tasks are made more complex and when measurements are taken

Situation encountered	Measurements
Start with 1 UAV	SAGAT
Add 2nd UAV	SAGAT
Add 3rd UAV	SAGAT
Induce errors in simulation	SAGAT
Add 4th UAV	SAGAT
Induce errors in simulation	SAGAT
End of test	NASA TLX

3 Implementation

Each of the created profiles is representative of a control task as defined before (Controlling a single UAV, controlling multiple UAVs, operating an onboard

camera and operating as a tactical commander). The improvements that were made are described in the following sub-sections while a direct comparison is shown in Fig. 1 and Fig 2.

3.1 PFD

One drastic change that was made was the removal of a classical primary flight display (PFD) present in all modern aircraft. Normally, such a PFD includes the same information as the basic T (airspeed indicator, attitude indicator, altimeter and heading indicator) [6].

There are several reasons for this step. First, heading information is already included in the main map. Second, the operator survey has shown that attitude information was not deemed critical. This is backed by the fact that the UAVs are not controlled directly but through a series of waypoints which are followed by the autopilot.

Instead of having a traditional PFD, the airspeed indicator and altimeter are coupled with the map. This has the advantage that operators need not deviate their focus from the map to assess the UAV's state. Additionally, this step increases consistency between single UAV and multi UAV display configurations. It is known that poor visual momentum - a concept borrowed from the film industry [7] - induces cognitive difficulties when switching between displays [8] [9]. So in order to improve the quality of the overall console, individual items may have to be designed in a non-optimal way [10].

3.2 Status Indicators

It is necessary for the operator to quickly detect any malfunctions the UAV might have. Tasks requiring integration of information rather than precise measurements are best served by object like displays [11]. Therefore, the text list of subsystem statuses currently present in Neptus is replaced by a set of indicators. These indicators show a green light when a subsystem is functioning correctly and change color to inform the operator of a failure. This means that operators can immediately detect any changes of subsystem statuses.

In order to provide a fast overview of multiple UAVs, all subsystems are aggregated in a single indicator when the operator is controlling multiple UAVs. This way operators only have to sample very few indicators to acquire the status of all UAVs.

3.3 Airspeed Indicator and Altimeter

There has been extensive research about how to present altitude and airspeed information to a pilot. The principal of pictorial realism [12] dictates that the indicator representation should match the pilot's mental model. This includes the differentiation between digital and analogue information, as well as the orientation (up and down) and shape (circular vs. linear). Displaying digital information that must be transformed to a mental model means that processing time is increased [13]. Therefore, a ruler type display is used in modern aviation.

In contrast to full sized aircraft, the UAVs designed through the PITVANT project fly at low speeds and altitudes. This has the advantage that while showing the full range of possible airspeeds and altitudes, the resolution is still high enough to perceive small differences. As a result, the scale does not change and only the indicator itself moves. This means the direction of the indicator is equivalent to the pilot's mental model and also the principle of the moving part is satisfied [12] [14].

As these principles of compatibility - which are among the most important guidelines for display design [15] - are satisfied, definite improvements are expected.

3.4 C4I Specific Improvements

In addition to the improvements mentioned before, several other changes were made. These changes were specific to the use of Neptus as platform. Among others, they include additional filtering of waypoints and vehicles to be displayed.

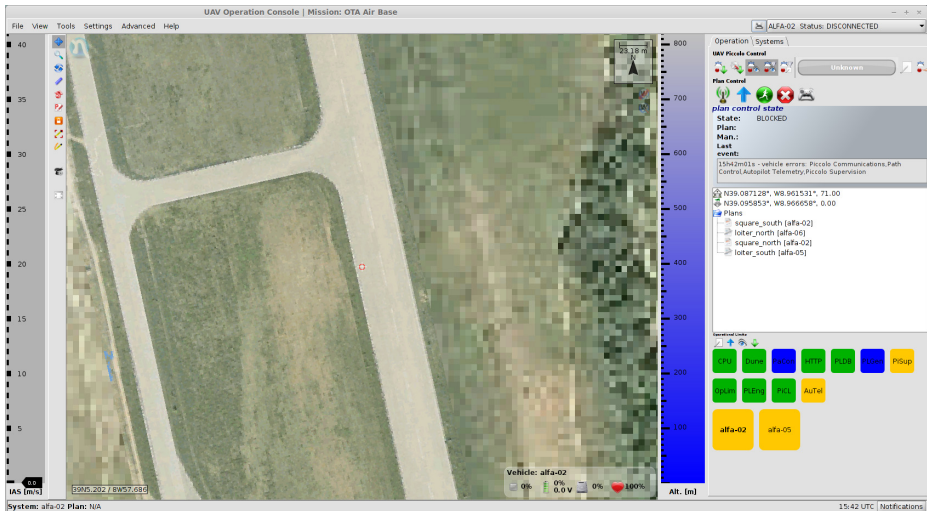


Fig. 1. Final console profile for simultaneous control of multiple UAVs

4 Results

The test was done with a total of 6 participants from the LSTS and the Portuguese Air Force Academy, comprising certified and uncertified UAV operators. The initial reaction of all participants was that the workload was too high and much higher than in a real operational scenario, which was as expected. Nevertheless, overall feedback was that the console profiles made the tasks significantly easier to cope with.

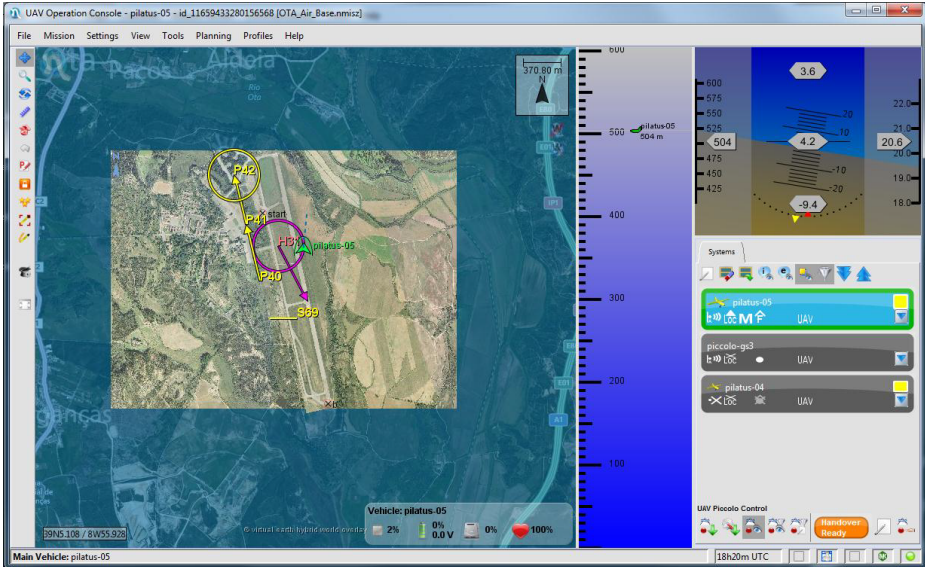


Fig. 2. Original console profile

Alongside these statements we have gathered test results. As can be seen in Table 3 and in Fig. 3, the average workload is reduced from 89.72 to 72.17, which is a reduction of 19.57 %.

Similarly, Table 4 and Fig. 4 show us that the average of correct answers increases from 51.62 % to 65.65 %, which is an increase of 27.17 %. The highest increase is shown for questions 4 and 12. It is noteworthy that for question 3, the percentage of correct answers actually drops.

Table 3. Total workload as measured with NASA TLX for each participant and console

Participant	Old console	New console	Reduction
1	81.00	73.33	9.47 %
2	92.33	63.00	31.77 %
3	87.67	81.00	7.60 %
4	94.67	64.67	31.69 %
5	96.67	76.67	20.69 %
6	86.00	74.33	13.57 %
Average	89.72	72.17	19.57 %

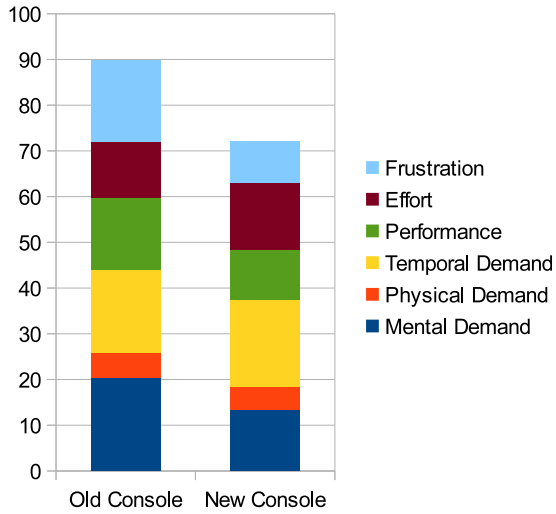


Fig. 3. Average workload as measured with NASA TLX

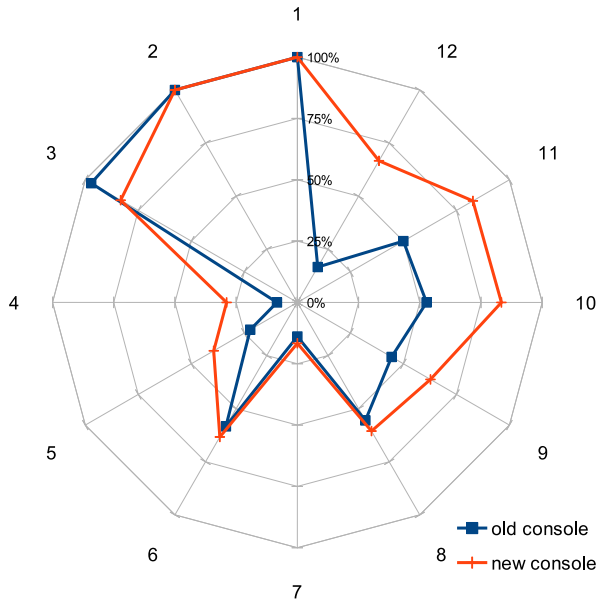


Fig. 4. Percentage of correct answers as measured with SAGAT

Table 4. Percentage of correct answers per question as measured with SAGAT for each console

Question	Old console	New console	Difference
1	100.00 %	100.00 %	0.00 %
2	100.00 %	100.00 %	0.00 %
3	97.22 %	83.33 %	-14.78 %
4	8.33 %	28.89 %	242.86 %
5	22.22 %	39.44 %	80.00 %
6	58.33 %	63.33 %	7.76 %
7	13.89 %	16.67 %	23.43 %
8	55.56 %	60.56 %	8.00 %
9	44.44 %	62.78 %	41.43 %
10	52.78 %	83.33 %	56.99 %
11	50.00 %	82.78 %	65.71 %
12	16.67 %	66.67 %	300.00 %
Average	51.62 %	65.65 %	27.17 %

5 Discussion

The results presented in Sect. 4 show a clear improvement in workload and situational awareness when using the new console profiles. In terms of situational awareness, 3 individual results stand out: Considerably higher improvement for determining all altitudes; Improvement in detecting anomalies; Deterioration of determining the main vehicle. Questions 4, 12 and 3, respectively.

The high improvement for determining all altitudes can be traced to the way that altitudes are presented. In contrast to the original console profile, the new profile dedicated to controlling multiple UAVs shows all UAV altitudes in the same indicator. This gives the operator constant access to that information without any switching of vehicles. While the number of correct answers for this question is still not very high, it should be noted that most operators could at least indicate the UAVs vertical separation with the help of the new console.

Similarly, the improvement in spotting anomalies (changed altitudes, airspeeds, communication disruptions, etc.) can be awarded to the newly added state indicator. This information was previously hidden and had to be actively sought for. Now it is prominently displayed, which attracts the operator's attention to any problem.

However we cannot ignore the deterioration detected when answering question 3. We believe that this can be traced to the fact that with the new capacity of observing all vehicles simultaneously, the operator loses sight of which vehicle he is currently issuing orders to. This trade-off forces us to re-evaluate the way that we currently present the main active vehicle.

6 Conclusion

In order to improve operator situational awareness and reduce workload, through information presentation control, alterations were made to a pre-existing operational C4I application. Moreover, feedback was gathered from certified UAV operators before development began and 4 different console profiles were crafted. Each of these profiles includes several improvements in terms of layout and display design. With the completion of these new profiles test sessions were held, in a simulated environment, with both certified and uncertified UAV operators. These tests showed that the average workload was reduced by 19.57 % while the situational awareness was improved by 27.17 %.

In summary, our initial hypothesis that changes made to the information's layout, and to the manner in which it is conveyed to the human operator, provide us a tool with which to control operator workload and awareness is supported by preliminary software in the loop tests.

Further Development. Although these results are promising, further tests are advised. Firstly, tests including real UAVs will provide more realistic stress levels and therefore provide a better workload gauge. Secondly, the operation scenarios must expand to include the other two profiles developed (video operation and tactical commander). Thirdly, the actual process of switching between profiles should be tested.

Finally, the different console profiles should be classified according to their levels of autonomy so that the process of switching between profiles can be automated [16].

References

1. Pinto, J., Calado, P., Braga, J., Dias, P.: Implementation of a control architecture for networked vehicle systems. In: IFAC Workshop on Navigation, Guidance and Control of Underwater Vehicles (NGCUV 2012), 270180 (April 2012)
2. Dias, P., Gomes, R., Pinto, J.: Mission planning and specification in the Neptune framework. In: Proceedings IEEE International Conference on Robotics and Automation, ICRA 2006, pp. 3220–3225. IEEE (2006)
3. Martins, R., Dias, P.S., Marques, E.R.B., Pinto, J., Sousa, J.B., Pereira, F.L.: IMC: A communication protocol for networked vehicles and sensors. In: OCEANS 2009-EUROPE, pp. 1–6. IEEE (May 2009)
4. NASA: Nasa task load index (tlx) (November 2012), <http://humansystems.arc.nasa.gov/groups/TLX/downloads/TLX.pdf>
5. Endsley, M.R.: Direct measurement of situation awareness: validity and use of SAGAT. In: Endsley, M.R., Garland, D.J. (eds.) Situation Awareness Analysis and Measurement. Lawrence Erlbaum, Mahwah (2000)
6. U.S. Department of Transportation, Federal Aviation Administration, F.S.S.: Instrument flying handbook (December 2012), <http://www.faa.gov/library/manuals/aviation/media/FAA-H-8083-15B.pdf>
7. Hochberg, J., Brooks, V.: The perception of motion pictures. Handbook of Perception 10, 259–304 (1978)

8. Wise, J.A., Debons, A.: Principles of film editing and display system design. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 31(1), pp. 121–124 (1987)
9. Woods, D.D.: Visual momentum: a concept to improve the cognitive coupling of person and computer. *International Journal of Man-Machine Studies* 21(3), 229–244 (1984)
10. Wickens, C., Andre, A., Haskell, I.: Compatibility and consistency in crew station design. In: Proceedings of the Ergonomics Society 1990 Annual Conference, pp. 118–122 (1990)
11. Melody Carswell, C., Wickens, C.D.: Information integration and the object display an interaction of task demands and display superiority. *Ergonomics* 30(3), 511–527 (1987)
12. Roscoe, S.N.: Airborne displays for flight and navigation. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 10(4), 321–332 (1968)
13. Grether, W.F.: Instrument reading. i. the design of long-scale indicators for speed and accuracy of quantitative readings. *Journal of Applied Psychology* 33(4), 363–372 (1949)
14. Roscoe, S.N., Corl, L., Jensen, R.S.: Flight display dynamics revisited. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 23(3), 341–353 (1981)
15. Wickens, C.D.: *Engineering Psychology and Human Performance* (1992)
16. Gonçalves, R., Ferreira, S., Pinto, J., Sousa, J., Gonçalves, G.: Authority sharing in mixed initiative control of multiple uninhabited aerial vehicles. In: Harris, D. (ed.) *Engin. Psychol. and Cog. Ergonomics, HCII 2011. LNCS (LNAI)*, vol. 6781, pp. 530–539. Springer, Heidelberg (2011)

A Web-Based Interface for a System That Designs Sensor Networks

Lawrence J. Henschen and J.C. Lee

Northwestern University, Evanston, IL 60208 USA

henschen@eecs.northwestern.edu, julialee@agep.northwestern.edu

Abstract. We describe the approach taken in the design of the interface for a system that helps application engineers who are not trained in computer science/engineering to design sensor networks. We cite various taxonomies from the sensor network literature that guided the design of the interface. We then describe the overall structure of the system to set the context for how the human interacts with it. We present some examples of the kind of data required to design a sensor network and describe how our interface collects that information. We note at many points in the presentation that a deep understanding of the data of the application allows for the design of an appropriate interface.

Keywords: Sensor networks, automated design, HCI.

1 Introduction

Sensor networks, both wired and wireless are rapidly becoming pervasive in modern society [1]. Applications range from monitoring patients [2] to agriculture [3] to structural health of buildings and other structures [4] and a multitude of other applications. A growing number of aspects of modern living involve some kind of collection and use of sensed data. Therefore it is important to make the design of sensor networks as simple and effective as possible. The design process requires knowledge and expertise in both the application (e.g., medicine or civil engineering) and computer engineering, the latter addressing the issues of sensors, computer nodes, networking, programming, etc. Thus, sensor network application projects currently require computer engineering specialists to select the computing hardware, design the network, and program the software running on the hardware. It would be much less expensive and more effective if the application experts themselves could specify the requirements for the sensor network in their own terms and have the software and hardware designed automatically. We believe this is achievable for two major reasons. First, the information required to design a sensor network is highly organized and well understood, as illustrated by the taxonomies described in the next section. Second, the hardware itself, such as sensors and computing nodes, is relatively simple compared to general computing hardware. While the second point makes automated systems of the type we are developing possible, the collection of required information about individual sensor network applications is crucial and requires a well-designed human-computer interface.

A key issue in making a system like ours usable to non-computer specialists is the design of the user interface. In this paper we show how a careful organization and analysis of sensor network applications and the information required to describe them led to the design of our user interface. Because the purpose of the interface was limited to collection of information, we could focus on a limited subset of general HCI principles and guidelines, as described for example in [5], such as learnability and selected guidelines from Schneiderman's Eight Golden Rules and Norman's Seven Principles. In particular, as will be discussed in later sections, the nature of the information required from the user led to a simple, mostly menu-driven interface that has high usability, high learnability, consistency, immediate feedback, good dialog control and closure, easy reversal of actions, and low memory load for the user.

This work contributes to the HCI field by illustrating that the analysis of the data involved in certain HCI applications can be used as the guide for applying general HCI principles, as does related work at Northwestern University on information display (as opposed to collection) [4]. We note here that our analysis was aimed more towards the content of the interaction than the visual format. Once the content, in our case the sets of questions to be posed to the user, is determined, it is straightforward to design good visual layouts. This is in contrast to other work at this conference, including our own work [4], in which the design of the visual aspects of the interface were a significant challenge even after knowing the salient characteristics of the data. Our work contributes to computer science in two ways. First, it opens the door to similar systems designed to help non-technical people design computer programs in general. Second, it provides a foundation and framework for the implementation of a fully automated sensor-network design system.

In Section 2 we give some background on sensor networks, particularly a variety of taxonomies for sensor network applications from which we derive a list of the information that is needed from sensor network designers. In Section 3 we describe the basic structure of the automated design system we are building. In Section 4 we describe some of the main features of user interface and how they are derived from a detailed analysis of the various taxonomies described in Section 2. Finally, in Section 5 we make some concluding remarks about the user interface design process.

2 Background - Sensor Networks

Sensor networks consist minimally of a set of sensor nodes and a base station, all of which communicate with each other, either wired or wirelessly. Each sensor node typically has a micro-computer and a set of sensors. The computer must be programmed to read the sensors at various times, collect the data, possibly perform some standard computations on that data, and transmit the data (either to other nodes or to the base station) at various times. In addition, the node may be programmed to accept commands from the network (other nodes or base station), coordinate with other nodes in the network, and a limited variety of other activities. Programming issues include, among others, scheduling sensor readings, transmitting data, controlling wake/sleep modes, and invoking data aggregation functions. Sensor node distribution or topology of sensor nodes in the application area is another important high level design issue.

Collecting the information necessary to automatically design such a system from non-computer specialists requires an interface that (1) allows the user to enter information in terms used by the application and (2) provides guidance from the system as to what kinds of information are needed. Designing an interface suitable for collecting the information requires an understanding of sensor network applications from many different viewpoints, especially if the system is to guide a user who is not a computer engineer but who must provide information needed for the design of the computing system.

Guiding the user through the collection of information requires a comprehensive understanding of a variety of aspects that affect sensor network design decisions. Fortunately, there is a collection of literature that provides various taxonomies that can aid in the design of an appropriate user interface. We give a brief summary of the most relevant taxonomies here.

- Bai et.al. [6] has provided a concise characterization of sensor network applications that covers essentially all the applications described in the literature at the time of publication. The characterization is based on the following eight properties.
 - Mobility - Do the sensor nodes move or not?
 - Sampling - Is data sensed periodically, continuously, only when an event happens, etc.? Does the sampling behavior change depending on events that happen, possibly in the node itself or in the network in general?
 - Transmission - Is data transmitted periodically, continuously, only when an event happens, etc.? Does the transmission behavior change depending on events?
 - Actuation - Does the node control some devices (machines, lights, etc.) outside the node?
 - Interaction - Does the node interact with other nodes in the network?
 - Data interpretation - Does the node interpret the data or simply collect and transmit it?
 - Data aggregation - Does the node aggregate data (e.g., take maxima, averages, etc.)?
 - Homogeneity - Are all the nodes in the network the same?
- Romer et. al. [7] describes the sensor network design space as a 14-dimensional space. Some of the dimensions address similar issues as Bai et. al. [6], for example: mobility, heterogeneity, connectivity. Other dimensions address other issues such as size, coverage, topology, lifetime, etc.
- Mottola et. al. [8] categorizes sensor network applications according to goal, interaction pattern, mobility, space, and time.

Many of the features have simple yes-or-no answers. For example, the nodes in the network are either all the same or not. Some have little or no effect on the general user interface design. For example, if the nodes are not homogeneous, then the user simply provides a description for each of the various kinds of nodes. Others features have wider ranges of values, but fortunately in most cases the range is small and well-defined. For example, sensor network nodes that do data aggregation almost always use maximum, minimum, average, etc. In some applications the sampled data

is interpreted rather than just aggregated. Interpretation algorithms are typically more complex than simple aggregation. Fortunately research in this area (e.g., [9]) provides knowledge that our system can apply. Our system allows the user to specify other aggregation functions and data interpretation functions and provide the corresponding code. Even for sampling behavior, the range of possibilities is limited - typically either periodic or based on some event. Most of the features listed in the various taxonomies impact one or only a few of the aspects of a sensor network. For example, if nodes are mobile then the range of mobility (feet, miles, etc.) would have an impact on physical aspects of the network, like power required for radio transmission; the only aspect relevant for programming is whether or not position is one of the kinds of data to be collected, transmitted, computed in an aggregate, etc. Conversely, whether or not an individual node computes aggregates is independent of how often the data is collected and would have almost no impact on the design of the communications protocol or the selection of radio transmitters.

It is important to note the distinction between collecting information about the requirements for the network and using that information to design an actual sensor network. We note that designing a sensor network from the ensemble of information collected from an application engineer is still a difficult problem and requires significant computation combined with extensive knowledge. How this is done is a topic for a computer science/engineering paper. Despite the complexity of the actual network design process and the large variety of types of information required from the user, the taxonomies cited here lead to a highly organized set of questions that are presented to the user and to which the user can provide quite simple answers. That is, the interface can be simple and straightforward even though the way the data are used is complex.

3 System Structure

Based on our study of the various taxonomies describing different aspects of sensor networks and our knowledge of computer science, in particular compiler theory and network theory, we have chosen a four-layer structure for our automated sensor network design system.

1. Application Layer
2. Model Layer
3. System Layer
4. Implementation Layer

The Application Layer is the web-based system for collecting information from users. The amount of information required to design a whole sensor network is quite large, but the various taxonomies mentioned in the previous section allowed us to organize the collection of questions into a hierarchical structure. How the taxonomies mentioned in Section 2 led us to this organization is one of the main contributions of this paper and is discussed in Section 4.

The Model Layer contains our internal representation of the sensor network, in particular our representation of the sets of nodes and their functionalities (source, sink, sensing behavior, etc.) and the network topology and related issues (number of nodes, spatial distribution, communication protocols, etc.). The Model Layer also contains the system's event model. Because the Application Layer user interface is designed to allow the user to enter information in the application's terms, the Application Layer information will not be in a format suitable for analysis and automatic generation of sensor network designs. We are designing data formats and structures that will facilitate the automatic generation of code, network layouts, communications protocols, and other aspects required for implementing an actual sensor network.

The System Layer is a collection of service functions that can be drawn on once the choices about software have been made in the Model Layer. These functions are organized in libraries, each of which relates to one aspect of the overall software model, for example event handling functions, data collection functions, communications functions, etc.

The Implementation Layer is concerned with the lowest level objects of a sensor network - the hardware elements (processors, nodes, sensors, wireless transceivers, etc.) and the specific software that runs on the various computer platforms in the network.

The major software components of our systems translate between the various layers. In particular, from the HCI point of view, the translation between the top two layers transforms information expressed in user/application terms into an internal, system-oriented representation and follows the notion of "interaction framework and translation between components" [5]. Translations between the lower layers are of interest to computer engineers and scientists but not of direct interest to the HCI community; therefore, they will not be discussed here.

4 Designing the User Interface

Our user interface has three major goals regarding the collection of the information from the user - (1) to allow the user to see information and enter information in his or her own terms, (2) to intelligently guide and prompt the user to enter the right kinds of information, and (3) to make it easy for non-computer engineers to enter the required information. The analysis of the variety of taxonomies presented in the literature described in Section 2 and our knowledge of technical issues in the design of sensor network hardware and software has led us to a hierarchical organization of the questions to be presented to the user. The top level allows our system to understand the application being addressed by the user and to tailor the remainder of the interaction for that specific user. Lower levels get into more of the details of how the network and its individual nodes are supposed to behave. The questions at each level are guided by information collected at higher levels. In most cases our system can present a set of choices to the user rather than requiring the user to type text answers.

4.1 Questions for Collecting Static Information about the Application

The first level of questions is concerned with the application. The initial question simply asks the user to select the broad area from a list (drop-down menu). The list would include the applications described in the literature such as “industrial”, “military”, “medical”, “environmental”, etc. Based on the user selection our system will follow up with more detailed questions about that specific application area and will use language appropriate for that area. For example, environmental applications have specific issues not applicable to medical or industrial applications. So, if the user selected environmental as the general application area, our system would follow with questions about environmental issues: indoor or outdoor? air or plants or animals? if outdoor what is the size of the area to be monitored? etc. Each input helps configure the next layer of questions. Assuming the knowledge base about each application is sufficiently robust, users are intelligently guided through all the relevant questions, even ones that the user might not have thought of but that the system knows are relevant. Moreover, the questions are put in the context of the application, not in the technical context of nodes, networks, and computer/sensor hardware. For example, for an outdoor environmental sensor network distance is an important issue because of power requirements for long distance transmission; the environmental engineer should not have to know about the reason for concern, only have to know how far apart the nodes in the network are. This, of course, requires our system to contain knowledge about sensor-network issues in each of these application areas. However, our system is not a general natural language interface system, so these knowledge bases will be of only moderate size. Our system is a highly focused information collection system in which the issues for each application area are reasonably well understood, and so the questions to be presented to users can easily be gleaned from the literature. Finally, this portion of our system is flexible and easily extensible as new application areas are studied and existing ones become more sophisticated and better understood. In summary, the structure of the questions is a tree that is dynamically generated by the answers entered by the user and the knowledge base, as illustrated in Figure 1. Note that this is an organizational tree, not a search tree; therefore, the tree can be traversed in an order (breadth first, depth first, other order) most suitable for interacting with the user.

We note that many generic kinds of questions will be common across many applications. The primary example is what is being sampled – e.g., temperature, humidity, etc. Examples of these kinds of questions include “What is being sampled?” (e.g., temperature, humidity, vibration, etc.) and “What is the geographic area or volume for the application?” However, often the questions still need to be tailored to match the expectations of the user. For example, distance is a common issue in many applications. However, distance for a military battlefield application would undoubtedly be measured in miles, whereas distance in an indoor environmental application would be measured in terms of feet or yards. Distance for an outdoor environmental monitoring application might be either – perhaps feet or yards for monitoring individual crop fields in an agricultural application but miles in a seismic monitoring application. Our system uses the information from the first battery of questions to phrase how these common questions are posed to the user.

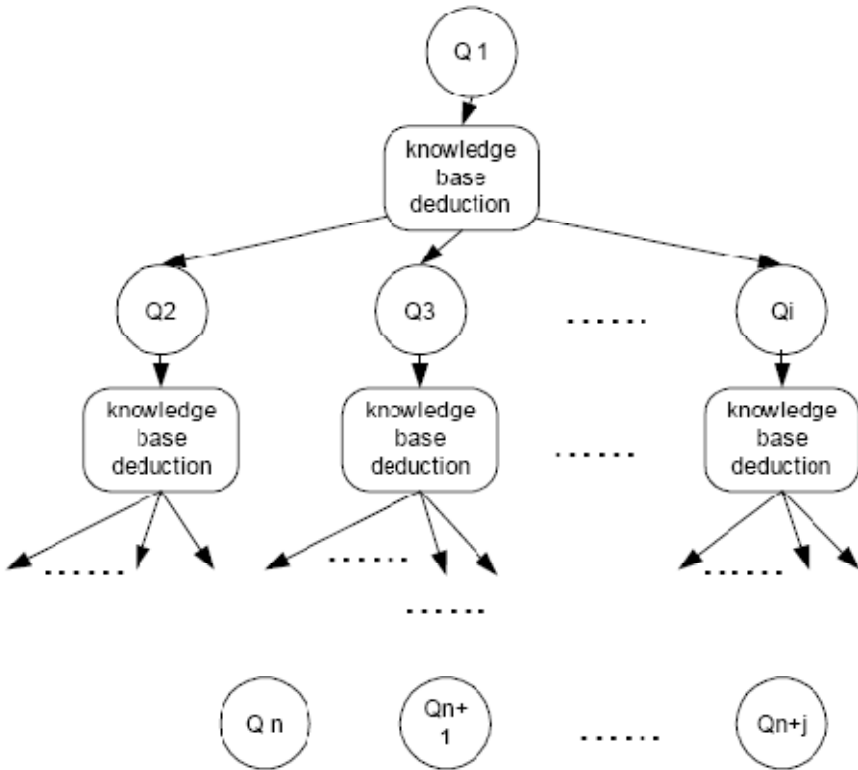


Fig. 1. Hierarchical Question Tree

4.2 Questions for Collecting Dynamic Information about the Application

After the user has entered the static information about the application itself, the next major set of questions collects information about the behavior of the sensors and individual sensor nodes. The first set of questions asks the user about the sampling behavior of each sampled item and the transmission behavior. For each kind of data to be sampled, there are two issues. First, our system will prompt for the sampling behavior, asking such question as how often should samples be taken (milliseconds, seconds, minutes, hours, etc.), whether or not sampling is based on events, under what conditions sampling behavior changes (see the next paragraph about events), and whether the node should store raw or aggregated data. Our system also asks whether or not sampling of several sensors should be coordinated. Second, our system will prompt for the transmission behavior, asking questions similar to the ones for sampling - how often are data transmitted to the network, does that behavior change based on events, should raw or aggregated data be transmitted. Events may change these behaviors; therefore, the user is prompted at this stage to indicate how many different behaviors there will be and to name them so that the event interface

(described in the next paragraph) can refer to them. Except for the entering of specific times through text boxes, the user can provide most of this information through drop down menu choices, which makes the interface easy to use and allows rapid entry of the necessary information.

Another set of questions collect information about the various kinds of events that can affect the behavior of the network. Some events affect behavior of individual nodes; examples of these include sensed values or combinations of values go out of range (e.g., temperature-humidity index goes above 100), sensed values exhibit a significant change (e.g., onset of vibration on a bridge), and explicit commands or information are received from the network (e.g., command from the base station or information that a neighboring node has failed). Our analysis of the literature indicates that events based on sensed data can be described by a combination of relations among values (e.g., $x > y$), the passing of thresholds (x increases past 100), and simple trends (e.g., x begins to decrease or x has been decreasing for y amount of time). An example from agricultural monitoring is

(soil moisture $< .15$) & (temperature increases above 100) &
(humidity has been decreasing for one hour).

The event itself occurs exactly when the set of conditions first becomes true; that is, events correspond to points in time. Behaviors may change at these points. The user interface for events is much more involved than that for the sensors for two reasons. The first, and obvious, reason is that the expressions used to describe events are much more complex than the information about sensors. Second, and perhaps as important, non-computer science/engineering users may not even be aware of the possibilities. Our system might address the first of these by the use of an expression writer similar to those found in many programming systems. Our system can also suggest the use of intermediate variables to hold values computed from a combination of sensed data, such as THI (temperature-humidity index) or aggregates (such as average or maximum). Our system addresses the second complexity issue by providing tutoring and guidance throughout the process. The user is given a list of possible relationships and event types, some quite general and others perhaps special to the application area specified by the user in the initial interaction. As with the questions about the sensors our system uses information provided by the user in the early stages of the interaction combined with facts and rules in the knowledge base to intelligently guide the user through this phase of the design.

4.3 Samples of Other Question Groups

Space limitations preclude a complete description of all the question groups and interface features. We mention a few more here and give a few examples of questions to give the reader a better view of the scope of the system and to emphasize again that a careful analysis of the nature of the information to be collected leads to an appropriate organization of the information, which leads in turn to the design of an appropriate user interface.

- **Actuation.** What are the devices to be controlled? Based on the application area our system can suggest common ones; for example, for an industrial application there would be machines, lights, etc. Our system would also apply knowledge about those, such as whether they are digital or analog. In the case of analog, the system would prompt the user for the range and resolution of the output, issues an application engineer might not think of.
- **Dynamics.** Do nodes move, or are they stationary? If they move, what is the expected speed and range?
- **Geography.** What is the volume of the area? Based on the application our system will suggest relevant dimensions such as meters or miles. Is it two dimensional or three dimensional?
- **Life Span.** How long must the network continue to operate (days, months, years, etc.)? Are the nodes accessible for battery replacement? Again, a non-computer specialist might not consider this question, but our system would need to know about it in order to decide among various hardware elements (which ones take less power) and whether or not to suggest energy harvesting additions to the nodes.
- **Reliability.** How reliable does the data need to be, and how reliable must the transmission be.
- **Budget.**

As noted several times, a non-computer science/engineering user would probably not think of many of these issues. Our system will understand which ones are relevant as more information is collected from the user and will explain the nature of the information required and guide the user in entering it.

5 Conclusion

We have described a system for automatically designing sensor networks from application-oriented user input. In particular, we have described the design of the human interface of the system and how the various taxonomies about sensor networks led us to that design. Our design was guided by HCI principles, the core one of which is usability. We feel that the system we are designing/developing will meet the needs of application experts for designing a sensor network for their specific application without the need of deep involvement of and interaction with computer experts. We did not address details of how our system uses the information gathered to actually design a network; those details belong to computer science/engineering and will be presented in a future work in a different venue.

We believe that in applications such as ours - that is, applications in which the main function is the collection of data - the nature and content of the information is the most important aspect that guides the design of the interface. In our case, the sensor network literature revealed a hierarchical structure to the information. Choices of answers to some basic questions led to the need for and restrictions on successive questions. Also, the ensemble of information could be identified and classified ahead of time. This led to a hierarchical organization of the questions, which in turn led to the design of the interface itself. The required information in other applications might

have different structure, which would lead to a different organization of the user interface. The point is that it is the nature of the data and the way the user thinks of it that should guide the design of the interface. The first step in the interface design process should be to obtain a deep understanding of both those items.

References

1. Marwedel, P.: *Embedded System Design – Embedded Systems Foundations of Cyber-Physical Systems*, 2nd edn. Springer, New York (2010)
2. Akyildaz, I.F., Su, W., Sankarasubramanian, Y., Cayirci, E.: Wireless sensor networks: a survey. *Computer Networks* 38, 393–422 (2002)
3. Sikka, P., Corke, P., Valencia, P., Crossman, C., Swain, D., Bishop-Hurley, G.: Wireless adhoc sensor and actuator networks on the farm. In: *Proc. of the Int. Symp. on Information Processing in Sensor Networks*, pp. 492–499 (2006)
4. Kosnik, D., Henschen, L.: Design and Interface Considerations for Web-enabled Data Management in Civil Infrastructure Health Monitoring. In: *Proc. of the 15th HCI International Conference* (2013)
5. Dix, A., Finlay, J., Abowd, G., Beale, R.: *Human Computer Interaction*, 3rd edn. Pearson/Prentice Hall, New York (2004)
6. Bai, L.S., Dick, R.P., Dinda, P.A.: Archetype-Based Design: Sensor Network Programming for Application Experts, Not Just Programming Experts. In: *Proc. of the Int. Conf. on Information Processing in Sensor Networks*, pp. 85–96 (2009)
7. Romer, K., Mattern, F.: The Design Space of Wireless Sensor Networks. *IEEE Wireless Communications* 11, 54–61 (2004)
8. Mottola, L., Picco, G.P.: Programming Wireless Sensor Networks: Fundamental Concepts and State of the Art. *ACM Computing Surveys* 49, 19.1–19.51 (2011)
9. Camps-Valls, G., Gomez-Chova, L., Munoz-Mari, J., Rojo-Alvarez, J.L., Martinez-Ramon, M.: Kernel-based Framework for Multitemporal and Multisource Remote Sensing Data Classification and Change Detection. *IEEE Transactions on Geoscience and Remote Sensing* 46, 1822–1835 (2008)

An Interaction Concept for Public Displays and Mobile Devices in Public Transport

Romina Kühn, Diana Lemme, and Thomas Schlegel

TU Dresden - Junior Professorship in Software Engineering of Ubiquitous Systems, Germany
{romina.kuehn,diana.lemme,thomas.schlegel}@tu-dresden.de

Abstract. Public displays increasingly find their way into public space and offer a wide range of information to the user. Currently, most of these displays just represent information without the chance to explore them or interact with them. In general, by technical enhancements in this field, more and more possibilities of interaction are given in different domains. This work presents interaction opportunities between public displays and users with mobile devices in the field of public transport. As a basis for understanding the usage and benefits of public displays it is necessary to have a closer look at different types of displays in the public domain, too.

Keywords: Interaction concept, mobile interaction, public display, public transport.

1 Introduction

In public space different kinds of content are spread through different communication channels, e.g. on so-called public displays. While the usage of analog posters, static information leaflets or different signs primarily focuses on functional aspects, public displays represent context-sensitive information systems whose contents can be adjusted dynamically. These systems offer free access to digital information on public screens to the user. For this reason, public displays increasingly replace conventional information and advertising media. Depending on their use, public displays present their content with or without attracting attention. The range of applications reaches up from passive screens of information to more active self-service terminals. Therefore we present various possibilities of usage as well as a concrete example from public transport.

2 Conceptual Explanations

In literature, the terms public display and digital signage are often used synonymously. In general, digital signage is a networked (audio-) visual information system [1]. The content can be digitally created, managed, and played anywhere on a digital device. For this purpose, information is provided by people, but also by sensors and databases (Figure 1 information delivery). A content management system

(CMS) running on a server allows an organized information representation on displays. There is the possibility to control single monitors directly or combining several screens into logical groups (Figure 1 information presentation). The content can also be presented independently, to constantly communicate the relevant information (Figure 1 information assignment). While digital signage is a distributed system, a public display represents only an end device which shows digital content. Therefore, a public display can be seen as a part of digital signage.

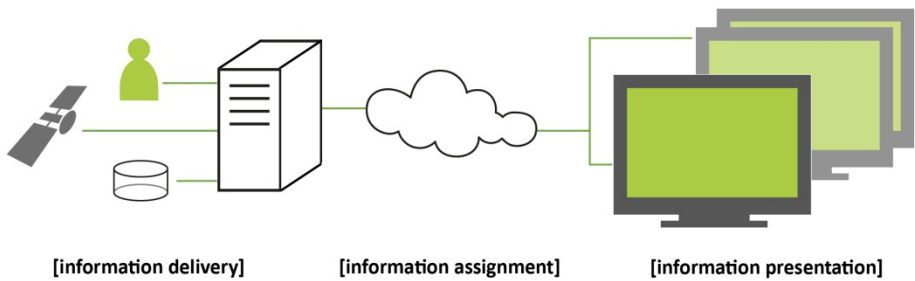


Fig. 1. Schematic construction of a public display system

To examine the interaction possibilities with public displays, different fields of application of public displays can be identified. Information screens or digital passenger information displays show the use of public displays as points of information (PoI) in a classical way. They represent the best solution to provide real time information to the target audience [2]. Typical installation areas are locations with high pedestrian traffic such as hotel lobbies, airports, train stations, department stores, shopping malls, self-service restaurants or museums [3]. In addition, kiosk systems and information terminals include users in an active way. Users can decide which information they want to receive, while they select information with offered input devices.

Points of sale (PoS) define places to purchase products or services - a physical location of a transaction [4]. The potential customers are pointed out to a product immediately. Ticket sales for public transportation such as trams, subways or railways or stamp-machines are just a few examples.

Another domain of public displays is the domain of so-called public playing. There are already some interactive games which allow users to play in public space. For example, a playing area is presented on a central screen and by entering an interaction area participants receive an invitation to play on their mobile device or by using different gestures.

In the field of public transport, information screens have been increasingly used since the first real-time information was installed in a metro station in Stockholm, Sweden in the 1980s [5]. It showed estimated waiting time in minutes and when a train was approaching, it also showed the destination of that particular train.

Self-service terminals in the form of ticket vending machines at airports and train stations are just a few examples for the usage of public displays nowadays.

The enhancement of possibilities of public displays as points of information is of particular interest for our work. Passengers consult displays at stopping points more often than printed information such as timetables [6]. Dynamic information displays in particular, are becoming more and more ubiquitous in modern public transport. Dziekan and Kottenhoff state that seventy to one hundred percent of people look at displays during waiting at a stopping point [7]. These displays show upcoming departures of trains, busses or trains at different stations leading to improved traveler information and service qualities.

3 Interaction Possibilities with Public Displays

For the mentioned kinds of application, interaction is necessary to enable the user to explore information. In case of the classical point of information, the display works as a simple information display. There is no interaction taking place according to human-computer interaction, since the user only acts as an observer who cannot participate in the event actively. Therefore we call this interaction a 1:0 interaction.

Apart from passive displays which just present information, we identified some interaction possibilities for public displays. A simple way to enable interaction is to combine displays with a keyboard, joystick or mouse. Due to technical enhancements in touch display technology this interaction option has increasingly found its way into public and semi-public spaces. It is an advantage, that these smartphone established touch interaction concepts can be used and so, the interaction between one display and one person (1:1 interaction) can be supported. However, strong signs of wear on the input device caused by dirt or vandalism make the usage of touch interaction displays in public space difficult.

By coupling with tracking systems or sensors, e.g. the Microsoft Kinect, better interactions, such as gesture- and voice-based interaction, with public displays can be provided. Based on the complex structure and strong context dependence of such systems they currently find just little use in public space. However, there are already successfully tested developments with gesture-based screen interaction [8].

Single or multi-touch screens as well as gesture- and speech-based technologies enhance the range of interaction with public displays. While they allow the user to explore information, they sometimes also have some disadvantages. For example, touching the displays is unsanitary and impossible in case of huge displays located too high on walls. Another disadvantage is the complexity of gesture-based displays which makes them difficult to build and to handle. Displays controlled by voice are limited in the public space, too.

Therefore, the use of personal mobile devices seems to be an alternative for interacting with public displays. The mobile device has to be connected to a public display, e.g. via Bluetooth or WLAN. The interaction can be graphically or gesture-based, whereby the gesture-based interaction can be carried out directly or with the mobile phone. Using the sensors of the mobile device, the input can be recognized. The input includes single-touch as well as multi-touch. Common touch gestures are presented and described in the Touch Gesture Reference Guide [9].

Furthermore, movements, which can be interpreted as different actions, can be carried out with the mobile device such as the interaction with touch gestures. This requires the use of sensors which are integrated per default to most of the current mobile devices, e.g. the gyroscope sensor. For example, the current orientation of a presentation in portrait or landscape format can be changed with that – both on the screen of the mobile device itself and on the associated public display. The mobile device can also be used with the help of tilting gestures for navigating in documents or moving of objects [10]. By doing so, the device is tilted over two imaginary axes, touching each other in the center of the device. Other gestures contain the throw and fetching gesture. They trigger, indicated by a fetching movement with the mobile device or a sweeping movement towards the body, sending and receiving data. Moreover, the shaking gesture, which can be interpreted by changes of inclination executed consecutively, causes deleting or resetting data.

Partly graphical forms of interaction base on the gesture interaction on the mobile device. A possibility, to use the data input in form of the keyboard on the mobile device, is to fill out forms shown on the display or to create documents. The data which is announced on the display can be presented in an abstract way on the mobile device, which can be carried out over colors, shapes, position, or signs. Content can be summarized to a region and represented as a specific shape on the mobile device, as shown in Figure 4. This abstraction is also suitable for detailed views by a certain area being closely zoomed in.

The interaction with public displays via mobile devices offers some advantages for us. Multimodal interaction possibilities are provided to the user. Furthermore hygiene problems and extensive technical constructions are dropped by the use of personal mobile equipment, which makes them interesting for various applications, for example in public transport.

4 Mobile Interaction for Public Displays in Public Transport

Due to frequent changes of schedules, printed timetables require intense maintenance and manpower. Therefore, a low-maintenance public display seems appropriate. In addition to a cost reduction for transportation companies, useful services can be put into practice. Schedule notices offer various information which must be processed for the use on a display. This affects the entire route map, single route schedules, departure times of various routes, fares for tickets (reduced, normal, group, etc.) and alternative routing.

This variety of data can be represented by the display size in a general view. For more detailed information it requires a higher level of detail. This can be produced individually and does not depend on a fixed time interval in which the content is updated automatically. The passenger must be able to interact with the display. This interaction is realized on the mobile device of the passenger, whereby a mobile device is connected to a public display. The use of the personal mobile device also enables data transmissions to the mobile device and the secondary use of this information, for example for route planning.

The concept of mobile interaction with a public display was realized prototypically for a fictional transport service. The passengers get the information mentioned above as well as the name of the current stop, date and time. This type of representation is also described as an information mode in the following Figure 2.

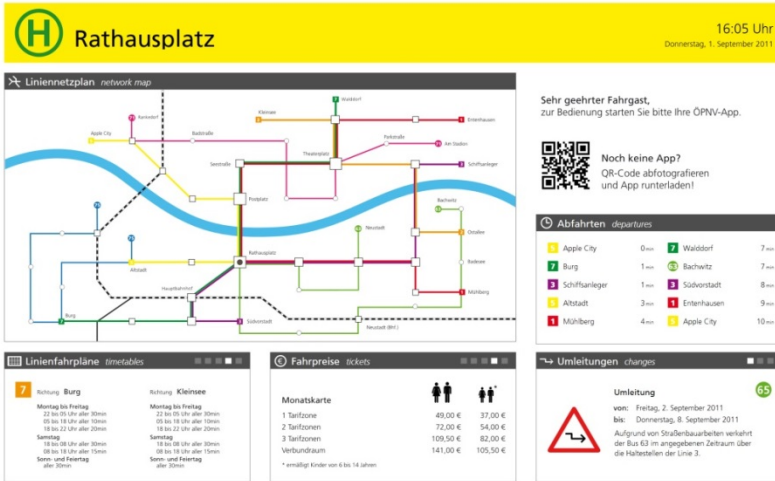


Fig. 2. Information mode of a Public Display

The user receives detailed information by connecting his mobile device to the display, which is performed by the user's mobile application. The display presentation changes to an interaction mode as shown in Figure 3.

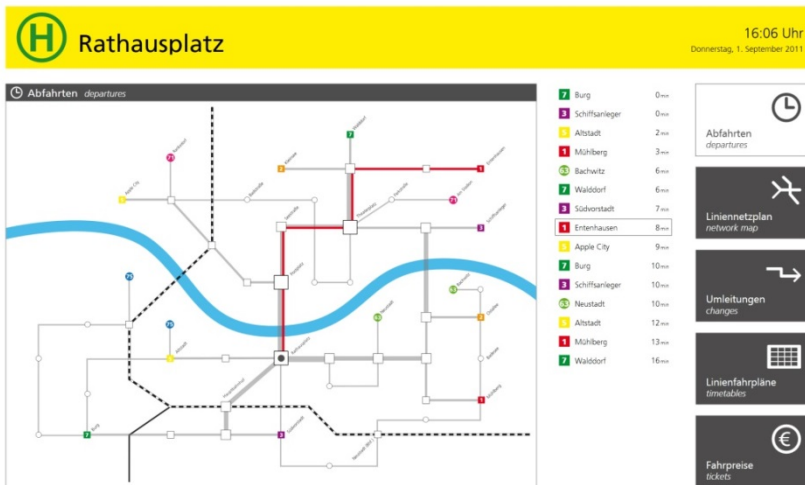


Fig. 3. Interaction mode of a Public Display

In the information mode the public display can be used completely for content representation. In the interaction mode the public display splits up into two areas – content on the left hand and navigation on the right hand (Figure 4). This splitting is also abstracted for the mobile device so that the user can interact with the mobile device as expected. The right area of the mobile device enables the user to navigate through the different kinds of content on the left hand side with help of different touch gestures. Different interactions are also possible with this concept, e.g. zooming in or moving content in the content area.

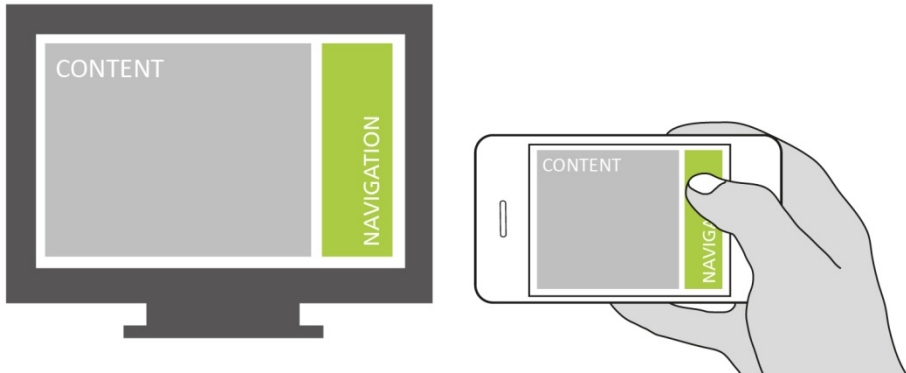


Fig. 4. Interaction concept for public display and mobile device

From the right navigation area, certain lines can be selected, information about alternative routes gathered, stops looked up, fares for certain routes calculated and tickets purchased. The information is presented in the left content area. In this way, the user is allowed to obtain details of separate areas of the journey planner individually and optionally transfer these to the personal mobile device. Therefore, the concept can be understood as a first step towards better interaction between mobile devices and public displays in the area of public transport.

This concept is also applicable to other domains. For public transport, the public displays can be used to simplify operation and maintenance of bus and tram stops, because, in conventional use, it is expensive and complicated to maintain every stop in a wide network. With displays information can be modified from a central point. Furthermore, the range of information can be more extensive and can include timetables, maps, tickets and different kinds of messages.

5 Conclusions and Future Work

To investigate the possibilities of interaction with public displays in public transport, different areas of application have been identified. They showed different forms of interaction with public displays, which are already used in public spaces, including the interaction with different input devices, such as touch, gestures and speech. The

interaction between personal mobile devices and public displays has been identified as a suitable way to obtain information in the field of public transport.

In this context we implemented a scenario to show the functionality and potentials of mobile interaction. Our prototype enables the user to interact with different kinds of information, e.g. timetables, network maps or tickets. Implemented as a base we can enhance our prototype with various opportunities, such as transferring important data to the mobile device.

Planned and shown as interaction between one display and one mobile device (1:1 interaction) the concept can be extended to more displays and more devices as well. This offers new possibilities in the field of public transport and the customer service. An example of 1:n interaction refers to passenger TV, which is already offered by many carriers. To obtain background information on individual topics, one or more users can use their mobile device to interact with the passenger TV. However, for n:m interaction across multiple displays and devices, an example is to use the interaction to find people joining a group ticket. Because of technical improvements of public displays and mobile devices, the interaction with these two technologies will become a more important service in public transport.

Another field of research is the adaption of public displays depending on the user's context. For example, important information can be highlighted on displays to help people finding their information. Furthermore, issues of privacy and questions about several users interacting with one display remain subjects of research. Sharing personal data on a public display was tested successfully by Peltonen et al. [11]. They developed a large public display where users can share their own media content using mobile devices. The so called CityWall was installed in a city center to show events happening in the city. In public transport, this kind of interaction and contextualization could be useful, especially to integrate traffic information such as delays or disruption.

Acknowledgements. We wish to thank Florian Schneider for the contributions made by his practical work.

References

1. GIM - Gesellschaft für Innovative Marktforschung, Digital Signage. Die globale Studie – Chancen und Risiken (2008)
2. Colavia, POI (Point of Information) with Digital Ads. Colavia Technology Co. Ltd. (February 18, 2013), <http://www.colavia.com/solution/poi.htm>
3. TVL Media Embedded Computing, Point of Information (POI) Kiosks (February 18, 2013), <http://www.tvlmedia.de/index.php/digital-signage-kiosks/poi-point-of-information>
4. Bars & Stripes, Point of Sale - A Beginners Guide to Computerized POS Software, The Small Business Depot (2005)
5. Schweiger, C.: Customer and media reactions to real-time bus arrival information systems (No. report 48), TRB (2003)

6. Dziekan, K., Sedin, S.: Customer reactions to the implementation of a trunk bus network in Stockholm. Paper presented at the 56th UITP World Congress (2005)
7. Vogel, D., Balakrishnan, R.: Interactive Public Ambient Displays: Transitioning from Implicit to Explicit, Public to Personal, Interaction with Multiple User. In: Proceedings of the 17th Annual ACM Symposium on User Interface Software and Technology (UIST 2004), pp. 137–146 (2004)
8. Villamor, C., Willis, D., Wroblewski, L.: Touch Gesture Reference Guide (2010), <http://www.lukew.com/ff/entry.asp?1071>
9. Dachselt, R., Buchholz, R.: Throw and Tilt - Seamless Interaction across Devices Using Mobile Phone Gestures. In: 2nd Workshop on Mobile and Embedded Interactive Systems (2008)
10. Peltonen, P., Salovaara, A., Jacuccia, G., Ilmonen, T., Arditob, C., Saarikko, P., Batra, V.: Extending Large-Scale Event Participation with User-Created Mobile Media on a Public Display. In: Proceedings of the 6th International Conference on Mobile and Ubiquitous Multimedia, pp. 131–138 (2007)

Study of Interaction Concepts in 3D Virtual Environment

Vera Oblaender and Maximilian Eibl

Chemnitz University of Technology, Chair Media Informatics, Chemnitz, Germany
{vera.oblaender,maximilian.eibl}@informatik.tu-chemnitz.de

Abstract. This paper describes what could be understood by interaction techniques and interaction concepts. In this work we focus in particular the second screen applications. Research of interaction techniques and concepts in this case investigates how to design interaction concepts with tablet as second screen, by remote connection with virtual environment on a primary screen. However, the actual samples used in this research are summarized by interactions like selection, manipulation and navigation aspects.

Keywords: Human computer interaction, second screen, manipulation, navigation on virtual environment in virtual reality, interaction technique, interaction concept, gestures.

1 Introduction

Interaction is a central term to the following study, especially the in the context of interaction concepts as well as interaction techniques. The focus is on interaction concepts between second screen as a remote control device and virtual environment (VE) that will be represented on a large primary screen. The next section gives a very brief introduction what belongs to the second screen. Furthermore, to show different interaction techniques and concepts and to apply this in the matter of second screen, that was the motivation. The distinction drawn in the literature between interaction technique and interaction concepts is not clear defined, so that we will show how to separate this.

There are currently no consistent standards in the field of interaction in 3D VE, as is the case when interacting with computers in 2D, working with windows, icons, menus and pointers interfaces. 3D interaction in a 3D space should facilitate the feeling of a complete immersion. That means that people who interacting with virtual reality to completely forget of their environment. Interactions accord of expectations of behavior.

Due to the additional third dimension, there is much greater variety of interaction techniques coupled with the different drafts of interaction concepts for selection, manipulation and object manipulation as soon as navigation in 3D VE. Therefore and as a first step, we focus on current interaction concepts, particularly dealing with object manipulation in 3D space using selected input devices.

Results from this reflection might yield to more or less established quasi-standards. The appropriate use of concepts are proven by detailed research and existing user

studies. The results provide suggestions for further scientific work. Are there revealing gaps with urgent need for further scientific research?

2 Related Work

Nowadays a broad variety of media has expanded, but on the constellations of end-use, too. Several years ago we use a mobile phone for only making calls, today is possible much more actions with just one device. Therefore, it is no surprise that in the use of mobile phones like tablets were combined or mixed with another media [1].

Second screen concept will be used in the context of social network, thereby to exchanging the information with friends of current TV program [2]. The second screen trend is from using of a mobile device for additional information of current program. A typical example of second screen functionality as a TV shown describes the FANFEEDS application [3]. A next one changing development is NextShare^{Mobile} an integrated second screen application for Apple iOS devices [4]. With NextShare^{Mobile} it is possible to wind a video forward and back from the comfort of your sofa using included remote control.



Fig. 1. Second screen in application

The concept of second screen we take up an issue and integrate this in a simulation system. A tablet device is the remote control for interactions on virtual environment. This realization system includes tasks to interact with construction of factory environment, definition of task what worker has to do through to simulation and valuation of simulated working process. In this work we analyze interaction concepts build on the example of NextShare^{Mobile}.

Second screen shall be limited to interact with both hands. The left hand hold on the device and only right hand can interact with the touchscreen. Therefore, we able to find out interaction concepts only for one hand or for a few fingers. In normal practice occurs interaction in virtual environments with virtual hand, space mouse, specific glasses, joystick or game controller (Wii, PlayStation). These examples of trackers are too expensive.

The high-performance smartphones have good sensors, accelerometer, such as digital compass, and gyro on the market today, so we can work with and use these mobile devices as an alternative by the said trackers. The interaction can be classified as selection, manipulation and navigation.

In search of definition term for interactive technique and interactive concepts no results were found. Here is what we think of. Interaction techniques are techniques describes how information with the input device, touch sequence or sensing enter into a computer. Under interaction technique is one part consists of hardware and part of software elements. For example virtual hand is one hardware technique and button on the screen is one software technique element. But what is about demarcation of interaction concept? The combination of different interaction techniques and specific chronology of interaction techniques with a device makes an interaction concept.

3 Interaction Concepts

The directed manipulation in 3D space is the most natural technique, because it is intuitive for people to act on physical objects. The second screen applies a semi-immersive interaction with a virtual environment because it is a remote interaction. But doesn't have to be necessarily a negative there are methods to obliterate the differences. This will happen by following concepts.

3.1 Manipulation of Objects with Remote Controller

Represented Example. A represented example for this research is a work of [5]. The focus of this work are to design "...on the 3D object manipulation similar to the traditional virtual hand researches". The communication between primary screen and mobile device is done over Wi-Fi. The placement of cursor on virtual environment is happened unanimously by moving of second screen device in all directions of axes. This describes continuous commands in the table 1. The interactions on the z axis doing by sweeping up and down the screen with a finger on the screen or by moving the mobile device by stretching the arm holding the device (object scaling).

Event-based command is for example to select an object. The placement of cursor in 3D VE should be for the object and then interact with finger by double tap (this is the grasp state). For rotation command you tap triple in grasp state. However, move device and the rotation on object in 3D is following. The conform behavior with the virtual object through device appeals in the hands for real object [6].

Table 1. Overview of the Commands. Source[5]

Continuous Command [Ⓟ]	
<i>3D Hand Placement</i> [Ⓟ]	Moving the virtual hand by moving the mobile device [Ⓟ]
<i>Object Translation</i> [Ⓟ]	Moving the object in the virtual hand by moving the mobile device [Ⓟ]
<i>Object Scaling</i> [Ⓟ]	Scaling the object by sweeping on the mobile device's screen [Ⓟ]
<i>Object Rotation</i> [Ⓟ]	Rotating the object by tilting the mobile device [Ⓟ]
Event-based Command [Ⓟ]	
<i>Grasp</i> [Ⓟ]	By double tap [Ⓟ]
<i>Release</i> [Ⓟ]	By tap [Ⓟ]
<i>Scaling Mode</i> [Ⓟ]	By double tap in grasp state [Ⓟ]
<i>Rotation Mode</i> [Ⓟ]	By triple tap in grasp state [Ⓟ]

Bimanual Surface. Here investigate another example [7]. This work shows an agreement between gestures for choice of sensory, multi-touch and dual-surface input with bimanual touch- and motionabled concept. There is a difference to the represented example. If an interaction is to translate the object that would be interesting for the z axis direction (Figure 2).

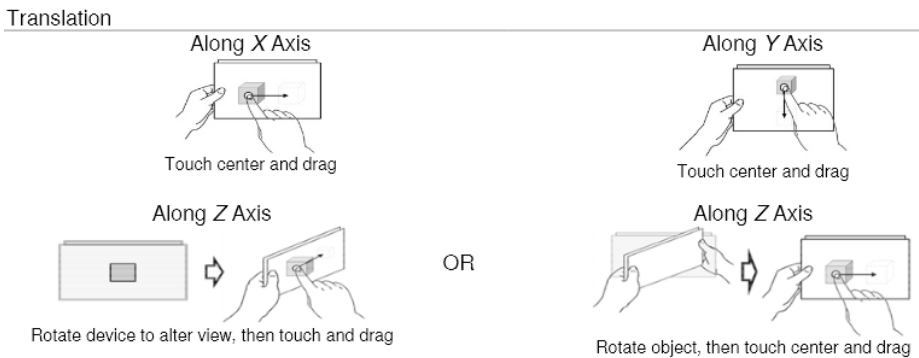


Fig. 2. Translation on z axis is a combination of gesture and a using device sensor [7]

The manipulation is happened direct on the touch screen. Gestures parameter should be transfer on primary screen. In this case we use a metaphor of 3D object from primary screen on second screen. The metaphor means an object on second screen that represent the object of VE, so can interactive with the object remote.

Definition of Detection Gestures

Another idea is define and create own detection gestures, one that describes interactions with finger. For example to select an object this will happen by five-finger-pinch [8]. It would be possible to define interactions for undo and redo operations. The problem of this detection is, there must be made unique association between gesture, object and the situation. Another aspect is the definition of new gestures needs to be learned. Is that attempt intuitive for everybody?

Points Tap

On the base of [9] [10] there are another attempt to interactive with mobile device. The interaction concepts are scheduled for two hand interaction, but this can transferred to one hand interaction. In which the points to determine on a metaphor object in second screen and then performed an interaction gesture with the finger (Figure 3). This approach to solve the interaction on z axis for 2D interactions without changing the direction or the perspective for interaction.

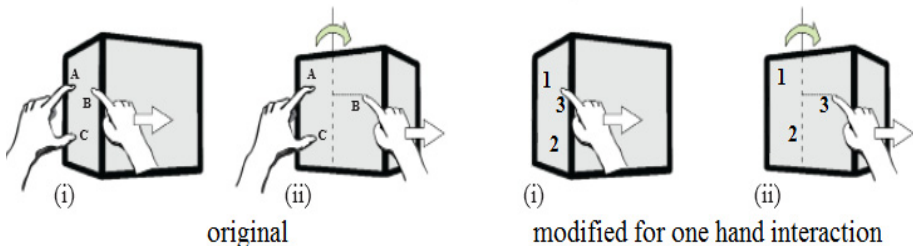


Fig. 3. On the left side is the original interaction and on the right side is the modified interaction for one hand [9].

Instead of A,C points for left hand like in original (i), can tap with the right hand on the points 1,2 and the third one is to move gesture (ii). Principal can so be manipulated all points in all axes in different situations.

This would be a selection of interaction concepts. In this case is a combination of the hardware techniques and sensors of tablet are for rude interactions and gestures techniques on touchscreen are for detailed indications from the manipulative object.

3.2 Discussion

For some of discussed concepts is to have a metaphor on second screen a requirement. However, here is another thing that should be considered. One problem stays, how to define an interaction for undo operation on mobile device?

4 Results

Is that a standard set for interaction techniques and interaction concepts in virtual environments/ virtual reality? Advantage of second screen application is the mobility of remote control. You can use the efficient sensors for more intuitive interaction and visualization of 3D EV is limited only by the screen size. Simulations and demonstration are showed by a large screen. In the other hand the disadvantages are to interact with one hand. Interaction would maybe only by 2D touchscreen and limited visualization on the second screen size.

5 Future Work

To design an interaction concept as a second screen application hosts some interaction restrictions. But this is the most interesting thing on development. We want to evaluate our user study of interactions with second screen application. To consider of an undo interaction concept is for future work, too.

What is a clear definition of interaction technique and interaction concept? Where are the differences? To design interactions with second screen interaction is a challenge.

References

1. Courtois, C., D'heer, E.: IBBT-MICT-Ghent University: Second Screen Applications and Tablet Users: Constellation, Awareness, Experience, and Interest. In: EuroITV, Berlin, Germany. ACM (2012)
2. Lochrie, M., Coulton, P.: School of Computing and Communications, Lancaster University: Mobile Phones as Second Screen for TV, enabling Inter-Audience Interaction. In: ACE, Lisbon, Portugal. ACM (2011)
3. Basapur, S., Mandalia, H., Chaysinh, S., Lee, Y., Venkitaraman, N., Metcalf, C.: Interactive Media User Research: FANFEEDS: Evaluation of Socially Generated Information Feed on Second Screen as a TV Show Companion. In: EuroITV, Berlin, Germany. ACM (2012)
4. Knowles, W., Mu, M., Bamford, W., Race, N., Needham, C.: Demo: Introducing NextShare^{Mobile}, An Interactive Second Screen Application. In: MobiSys, Low Wood Bay, Lake District, UK. ACM (2012)
5. Lee, D., Kim, G.J., Hwang, J.-I., Ahn, S.C.: 3D Interaction Using Mobile Device on 3D Environments with Large Screen. In: MobileCHI, Stockholm, Sweden. ACM (2012)
6. Daiber, F., Li, L., Krüger, A.: Designing Gestures for Mobile 3D Gaming. In: MUM, Ulm, Germany. ACM (2012)
7. Liang, H.-N., Williams, C., Semegen, M., Stuerzlinger, W., Irani, P.: User-defined Surface+Motion Gestures for 3D Manipulation of Objects at a Distance through a Mobile Device. In: APCHI, Matsue-city, Shimane, Japan. ACM (2012)
8. Lü, H., Li, Y.: Gesture Coder: A Tool for Programming Multi-Touch Gestures by Demonstration. In: CHI, Austin, Texas, USA. ACM (2012)
9. Reisman, J.L., Davidson, P.L., Han, J.Y.: A Screen-Space Formulation for 2D and 3D Direct Manipulation. In: UIST, Victoria, British Columbia, Canada. ACM (2009)
10. Bollensdorff, B., Hahne, U., Alexa, M.: TU Berlin: The Effect Perspective Projection in Multi-Touch 3D Interaction. In: Graphics Interface Conference Toronto, Ontario, Canada. ACM (2012)

Undo/Redo by Trajectory

Tatsuhito Oe, Buntarou Shizuki, and Jiro Tanaka

University of Tsukuba, Japan
{tatsuhito,shizuki,jiro}@iplab.cs.tsukuba.ac.jp

Abstract. We have developed a trajectory based undo/redo interface. Using the interface, a user traces actions' trajectories shown on a display. As a result, undo/redo manipulations are performed rapidly with selection of a target from a history. In this paper, we describe interaction techniques, implementation, and advanced usages of the interface.

Keywords: undo/redo, trajectory, history, tracing, desktop interface, gui.

1 Introduction

In many applications, user's actions (e.g., key commands or mouse actions) are stored as history for undo/redo (Figure 1a). When the user undoes/redoes these actions, the user executes commands for undo/redo one or more times (Figure 1b). Because the user must execute commands several times, performing undo/redo is time-consuming.

To perform undo/redo manipulation faster, the following approaches have been developed that allow the user to undo/redo actions selectively.

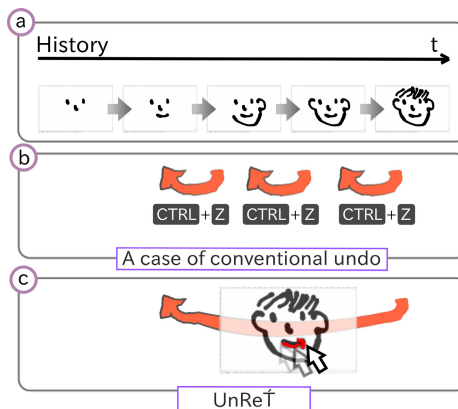


Fig. 1. Conventional undo and UnReT's undo in linear history model

Some approaches visualized a history using texts or screenshots. For example LibreOffice Impress¹ showed a list of manipulations as texts, and Meng et al. [13] presented a visualization of selective undo [6] using screenshots.

Other approaches localized undo/redo manipulation's region. For example, regional undo model [9] enables the user to undo/redo actions that occur in a specific region of a display. Moreover, a selective undo model [6] enables the user to undo/redo an isolated action.

These two approaches have an advantage and a disadvantage. The former approach enables a user to look an overview of a history rapidly, though takes a time to undo/redo if the history becomes large. By contrast, the latter enables the user to undo/redo rapidly even when the history is large, though it cannot be applied to the linear history model used in most applications.

Our goal is to explore the effectiveness of an undo/redo interface where a user can undo/redo by selecting a manipulation directly. To achieve our goal, we developed a trajectory based undo/redo interface (*UnReT*²). Using *UnReT*, the user can undo/redo by tracing a mouse trajectory (one trajectory consists of mouse press, drag, and release) as shown in Figure 1c. Because the user does not need to execute many commands, the user can perform undo/redo manipulation faster.

2 Related Work

UnReT is an application-independent interface implemented by extending a desktop environment in which a user undoes/redoes actions by tracing a past trajectory visualized on the desktop, even on an application employing the linear history model. Therefore, related research into the interface includes work on history visualization using texts or screenshots, a history model, and a desktop extension.

2.1 History Visualization Using Texts or Screenshots

Some research has tried to visualize a history using texts or screenshots. In this research area, simple visualization uses a list of texts or screenshots. For example, Meng et al. [13] visualized the history using a list of screenshots.

In contrast to the simple visualization, some research has tried to represent additional information over texts or screenshots. Kurlander et al. [12] visualized actions' context using pairs of two screenshots before and after an action. Nakamura et al. [15] overlaid GUI actions on a screenshot to improve search speed from the history. In addition, Vratislav [19] adopted Fisheye Menus [5] to the texts' list of the history to improve search speed.

In this research, a user undoes/redoes actions by viewing and selecting an action from the list. By contrast, in *UnReT*, the user undoes/redoes by tracing a trajectory shown on a display. Therefore, using *UnReT*, the user does not need to look at the list, so the user can undo/redo faster.

¹ <http://www.libreoffice.org/>

² *UnReT* is short for “**U**ndo/**R**edo by **T**rajectory”.

2.2 History Model

A history model has been researched that localizes undo/redo's manipulation region spatially.

Berlage [6] and Myers et al. [14] presented selective undo model that enables a user to undo/redo an isolated action. Kawasaki et al. [9] presented regional undo model that has a broader region for undo/redo manipulation than the selective undo model. These history models have features to localize manipulation's region spatially. Therefore, these models can be applied not only to a single user's environment but also to a multiple users' environment where one user performs undo/redo actions in his/her own region [17,18].

These history models cannot be applied to a general application based on the linear history model. On the other hand, UnReT can be applied to various applications based on the linear model, because UnReT's implementation is independent from a particular application. That is, the implementation is based on mouse manipulations' trajectories stored in a desktop environment.

2.3 Desktop Extension

UnReT is the interface that extends normal undo/redo manipulation by capturing mouse trajectory. Related to UnReT, there is research on extending history or mouse manipulation in a desktop environment.

Extending history. Interfaces have been researched that assist a user by capturing his/her actions or states in a desktop environment. For example, Rekimoto [16] showed Time-Machine Computing that can recover any past state in the desktop. In addition, Kelly et al. [10] presented Desktop History that captured actions to visualize files manipulated at any application. Furthermore, Grossman et al. [8] captured actions and videos in the desktop to allow the user to watch the videos at any past action.

In these interfaces, captured actions are used for presenting a user's past manipulations. By contrast, captured actions are used for undo/redo manipulation in UnReT.

Extending mouse manipulation. There has been research on extending a user's mouse manipulation in a desktop environment. Appert et al. [3] developed Dwell-and-Spring, which enables a user to undo and cancel mouse manipulation on the basis of a spring metaphor between a cursor and a target. Kobayashi et al. [11] presented Boomerang, which allows the user to move files between directories by using throwing gesture. Similar to the above research, UnReT extends undo/redo manipulation also by using mouse manipulation in the desktop.

3 Undo/Redo by Trajectory

UnReT is an interface where a user can undo/redo by tracing a past mouse trajectory. Below is how to undo/redo in UnReT.

Undo: The user holds down the Ctrl key and traces a trajectory with a pointer. As a result, past actions including the one that gave the trajectory are undone (Figure 2a).

Redo: Holding a Ctrl key for long enough visualizes possible target trajectories. Thereafter, the user traces a trajectory. As a result, redo is performed (Figure 2b).

To support these manipulations, the system provides the functions below.

Preview of undo/redo. As shown in Figure 2c, the user can perform *Scratch gesture* while tracing the trajectory. This shows preview screenshots of undo/redo targets. Thereafter the user can undo/redo by selecting one of the screenshots and also cancel undo/redo manipulation with the “Cancel” button shown in Figure 2c.

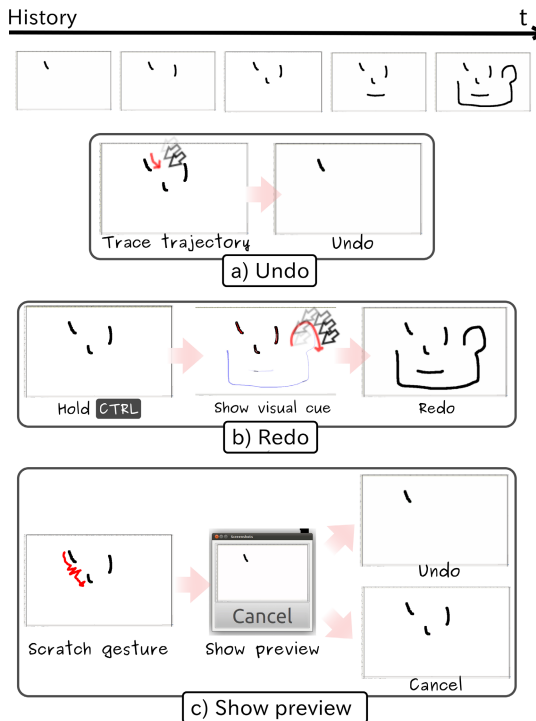


Fig. 2. Undo/redo interactions using UnReT

Trajectory visualization. When the user holds down the Ctrl key for long enough, mouse trajectories are visualized (Figure 3a). Trajectories are red for an undo target and blue for a redo target (Figure 3b). This visualization helps the user to know where to trace.

Showing a list of screenshots when tracing *similar trajectories*. Similar trajectories occur especially in contour drawing (Figure 4a). When the user traces such trajectories, applicable targets are shown as the list of screenshots (Figure 4b), and then the user selects an item from the list for performing undo/redo.

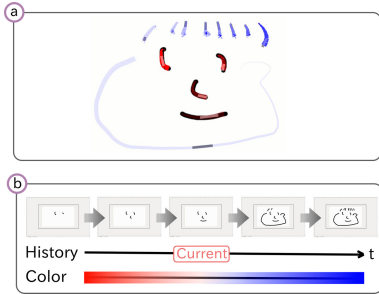


Fig. 3. Showing visual by holding down Ctrl key for long enough

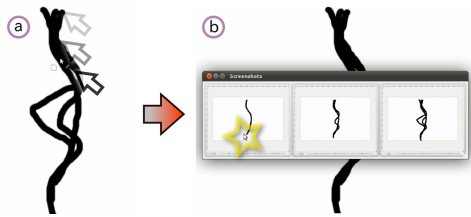


Fig. 4. Showing the list of screenshots when similar trajectories are traced

4 Implementation

Implementation of UnReT consists of mouse trajectory recording (Figure 5a), mouse trajectory matching (Figure 5b), and undo/redo processing (Figure 5c).

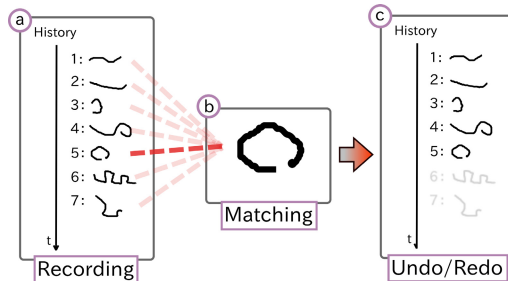


Fig. 5. UnReT consists of recoding, matching, and undo/redo processing

4.1 Mouse Trajectory Recording

When a trajectory is input, the trajectory's absolute points on a display are recorded to a history (Figure 6).

4.2 Mouse Trajectory Matching

When a user traces a trajectory while holding down the Ctrl key, similarities between the input trajectory and trajectories in the history are calculated using *Dynamic Programming (DP) Matching* (Figure 7). Then the system undoes actions until the trajectory with the highest similarity is input.

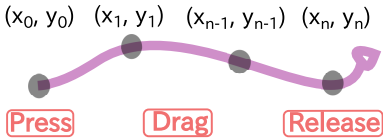


Fig. 6. Mouse trajectory recording

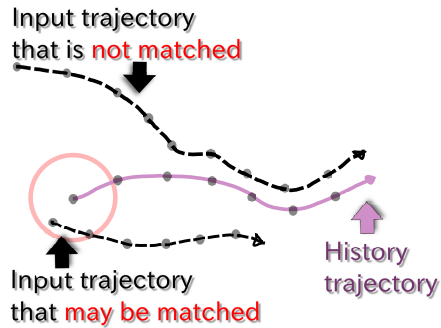


Fig. 7. Mouse trajectory matching

4.3 Undo/Redo Processing

In undo/redo processing, the system sends key shortcuts multiple times to the application. Because key shortcuts differ for each application, we implemented a function that enables a user to register undo/redo processing below.

undo/redo processing by key shortcuts. The system sends a registered key shortcut.

undo/redo processing by reversed mouse manipulation. Reversed mouse manipulation means inputting mouse's release, drag, and press from a trajectory in the history. In this process, Figure 6's (x_n, y_n) becomes the press point, $(x_{n-1}, y_{n-1}), \dots, (x_1, y_1)$ become drag points, and finally (x_0, y_0) becomes the release point.

Using the register function, a user can register a process such as "If an application is GIMP, the system sends Ctrl+z key event".

5 Applications

We applied UnReT to various environments and manipulations to explore UnReT's effectiveness. In this section, we describe trial results.

5.1 Environments

We tested UnReT in environments using a mouse and a keyboard, a stylus interface, and a touch interface, as shown in Figure 8. In every environment, UnReT was used by the first author using a GIMP application. Each environment’s trial results are shown below.

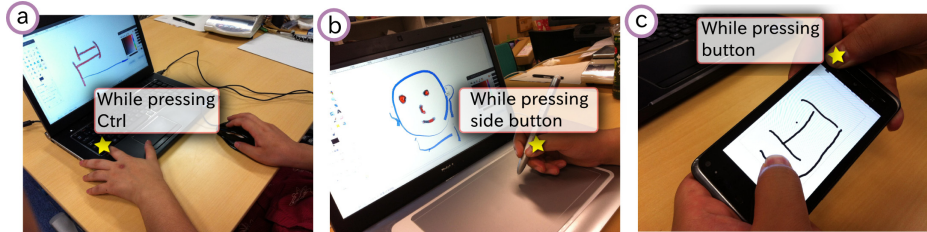


Fig. 8. Environments where UnReT was tested

Mouse and keyboard (Figure 8a). We used UnReT by using the mouse with the right hand while pressing the Ctrl key with the left hand (Figure 8a★). As a result, we observed that UnReT worked well in the environment. However, when the user needed to undo/redo manipulation only several times, inputting key shortcuts was more effective than UnReT.

Stylus interface (Figure 8b). We used UnReT with the stylus whose side button (Figure 8b★) was assigned as the Ctrl key. The first author did not use the stylus interface usually. Nevertheless, using the stylus made it easier to perform UnReT than using the mouse and the keyboard. The reason is that the stylus is a more suitable interface to trace a trajectory than the mouse. This result was consistent with that of Accot et al. [2], which revealed that the stylus performs better than the mouse for tracing by evaluating tracing performance of input devices based on Steering Law [1].

Touch interface (Figure 8c). We used UnReT on a mobile device. In this environment, we traced a trajectory with the left hand while holding down the mobile device’s button with the other hand (Figure 8c★). This environment was realized by accessing a remote Linux server running our system through a VNC client on the mobile device. As a result, using the touch interface enabled the user to trace the trajectory directly. This result could be applied to various touch interfaces.

5.2 Manipulations

We tested UnReT with various manipulations. Similar to Section 5.1, UnReT was used by the first author in every manipulation.

Drawing manipulation. We used UnReT for drawing manipulation using GIMP. As a result, we found a technique that enabled a user to perform

undo/redo manipulation before *changing a color*. The technique is performed as follows. At first, a user draws something (Figure 9a). Next, the user changes the color (Figure 9b) and then draws (Figure 9c). After that, the user can look at the trajectory (Figure 9d \star) when the user presses the Ctrl key for longer. Finally, the user can undo actions before changing the color by tracing the trajectory (Figure 9e).

We demonstrated drawing manipulation at a seminar of our university. At the seminar, we obtained comments such as “I want to undo/redo only an isolated action I trace”, “Is there any idea of how to apply UnReT to other history models?”.

For meet these requests, we need to adapt the selective undo model [6] to UnReT. To implement selective undo in UnReT, we plan to use the script undo model [4]. In the script undo model, if there is a sequence of actions (A_1, \dots, A_n) , a user can undo/redo an isolated action $(A_i(1 \leq i \leq n))$ like in the selective undo model. This is done by undoing A_n, \dots, A_{i+1}, A_i normally and then restoring A_{i+1}, \dots, A_n using a script. By restoring recorded trajectories as the script, we can implement selective undo in UnReT.

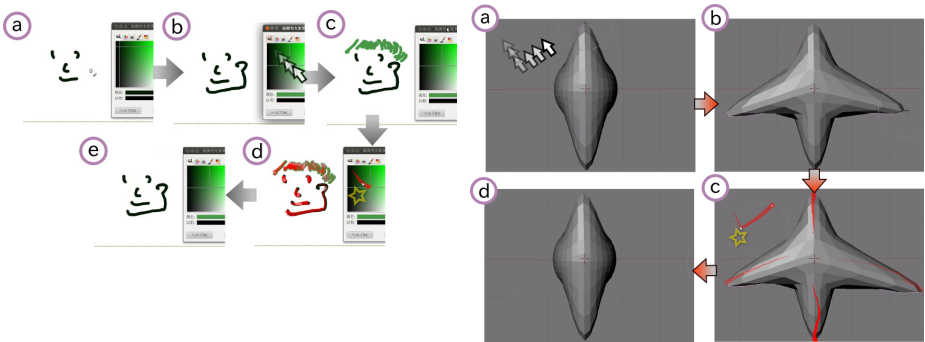


Fig. 9. Undo/redo before *changing a color*, **Fig. 10.** A user defines a mark for which is typically impossible in GIMP undo/redo

Marking manipulation. Marking manipulation enables a user to define shortcuts for GUI manipulations like UIMarks [7]. Figure 10 illustrates how to apply marking manipulation to Blender³ 3D modeler. In this example, first the user performs marking (Figure 10a). Then the user deforms the 3D model (Figure 10b). After that the user looks the trajectory of marking by holding down the Ctrl key for longer (Figure 10c \star). Finally, the user traces the trajectory and undo/redo the action until the marking (Figure 10d).

Icon manipulation. Using the reversed mouse manipulation as undo/redo processing, we enable a user to undo/redo icon movements, which is impossible in a typical desktop environment. In Figure 11, first the user moves the icon

³ <http://blender.jp/>

(Figure 11a). Then, the user looks at the icon movements' trajectories by pressing the Ctrl key for longer (Figure 11b). After that the user undoes icon movements by tracing the trajectory (Figure 11c). Similar to UnReT, icon movements were automated by Sikuli [20], using a programming by example of screenshots. In contrast to Sikuli, UnReT enabled the user to undo/redo icon movements.

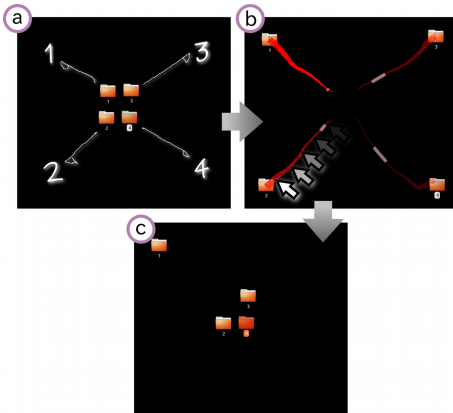


Fig. 11. Undo/redo icon movements

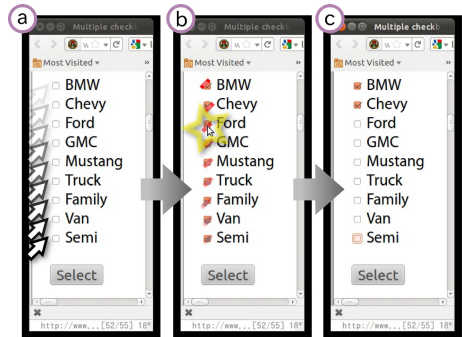


Fig. 12. Undo/redo checkbox selections

Checkbox manipulation. Similar to icon manipulation, a user can undo/redo selected checkboxes using the reversed mouse manipulation. In Figure 12, first the user selects checkboxes from top to bottom (Figure 12a). After that, the user clicks “Ford” (Figure 12b) while pressing the Ctrl key and then undoes selected checkboxes (Figure 12c).

6 Conclusion and Future Work

In this paper, we presented “Undo/Redo by Trajectory (UnReT)” and interaction techniques using UnReT. We applied UnReT to various environments and manipulations to explore effectiveness of the interface. Our future work is to implement selective undo manipulation in UnReT for further exploration.

References

1. Accot, J., Zhai, S.: Beyond Fitts' Law: Models for Trajectory-Based HCI Tasks. In: Proc. CHI EA 1997, pp. 250–250. ACM (1997)
2. Accot, J., Zhai, S.: Performance Evaluation of Input Devices in Trajectory-based Tasks: An Application of The Steering Law. In: Proc. CHI 1999, pp. 466–472. ACM (1999)
3. Appert, C., Chapuis, O., Pietriga, E.: Dwell-and-Spring: Undo for Direct Manipulation. In: Proc. CHI 2012, pp. 1957–1966. ACM (2012)

4. Archer Jr., J.E., Conway, R., Schneider, F.B.: User Recovery and Reversal in Interactive Systems. *ACM Transactions on Programming Languages and Systems (TOPLAS)* 6(1), 1–19 (1984)
5. Bederson, B.B.: Fisheye menus. In: *Proc. UIST 2000*, pp. 217–225. ACM (2000)
6. Berlage, T.: A Selective Undo Mechanism for Graphical User Interfaces Based on Command Objects. *ACM Transactions on Computer-Human Interaction* 1(3), 269–294 (1994)
7. Chapuis, O., Roussel, N.: UIMarks: Quick Graphical Interaction with Specific Targets. In: *Proc. UIST 2010*, pp. 173–182. ACM (2010)
8. Grossman, T., Matejka, J., Fitzmaurice, G.: Chronicle: Capture, Exploration, and Playback of Document Workflow Histories. In: *Proc. UIST 2010*, pp. 143–152. ACM (2010)
9. Kawasaki, Y., Igarashi, T.: Regional Undo for Spreadsheets. In: *Proc. UIST 2004 Demonstration Abstract*. ACM (2004)
10. Kelly, S.U., Davis, P.J.: Desktop History: Time-based Interaction Summaries to Restore Context and Improve Data Access. In: *Proc. INTERACT 2003*, pp. 204–211. IOS Press (2003)
11. Kobayashi, M., Igarashi, T.: Boomerang: Suspendable Drag-and-Drop Interactions Based on a Throw-and-Catch Metaphor. In: *Proc. UIST 2007*, pp. 187–190. ACM (2007)
12. Kurlander, D., Feiner, S.: A Visual Language for Browsing, Undoing, and Redoing Graphical Interface Commands. In: *Visual Languages and Visual Programming*, pp. 257–275. Plenum Press (1990)
13. Meng, C., Yasue, M., Imamiya, A., Mao, X.: Visualizing Histories for Selective Undo and Redo. In: *Proc. APCHI 1998*, pp. 459–464. IEEE (1998)
14. Myers, B.A., Mcdaniel, R.G., Miller, R.C., Ferrency, A.S., Faulring, A., Kyle, B.D., Mickish, A., Klimovitski, A., Doane, P.: The Amulet Environment: New Models for Effective User Interface Software Development. *IEEE Transactions on Software Engineering* 23(6), 347–365 (1997)
15. Nakamura, T., Igarashi, T.: An Application-Independent System for Visualizing User Operation History. In: *Proc. UIST 2008*, pp. 23–32. ACM (2008)
16. Rekimoto, J.: Time-Machine Computing: a Time-centric Approach for the Information Environment. In: *Proc. UIST 1999*, pp. 45–54. ACM (1999)
17. Seifried, T., Rendl, C., Haller, M., Scott, S.: Regional Undo/Redo Techniques for Large Interactive Surfaces. In: *Proc. CHI 2012*, pp. 2855–2864. ACM (2012)
18. Shao, B., Li, D., Gu, N.: An Algorithm for Selective Undo of Any Operation in Collaborative Applications. In: *Proc. GROUP 2010*, pp. 131–140. ACM (2010)
19. Vratislav, J.: Cascading undo control. In: *Bachelor Thesis*, pp. 1–52. Czech Technical University, Prague Faculty of Electrical Engineering (2008)
20. Yeh, T., Chang, T.H., Miller, R.C.: Sikuli: Using GUI Screenshots for Search and Automation. In: *Proc. UIST 2009*, pp. 183–192. ACM (2009)

Multi-layer Control and Graphical Feature Editing Using Server-Side Rendering on Ajax-GIS

Takeo Sakairi*, Takashi Tamada, Katsuyuki Kamei, and Yukio Goto

Advanced Technology R&D Center, Mitsubishi Electric Corporation
Sakairi.Takeo@db.MitsubishiElectric.co.jp,
{Tamada.Takashi, Kamei.Katsuyuki}@bx.MitsubishiElectric.co.jp,
Goto.Yukio@aj.MitsubishiElectric.co.jp

Abstract. This paper presents the methods of the multi-layer control and the graphical feature editing by the server side rendering on Ajax-GIS. Ajax-GIS uses divided raster image file called "tile" in order to keep light handling. We propose that the multi-layer control is realized by means of merging transparent tiled images in the server application as the requests of the client application. Furthermore we propose the graphical feature editing protocol that sent from a client and send back to an image in order to edit a feature such as moving vertices, changing color. In an evaluation experiment of an actual map data, we confirmed the effectiveness of these methods as compared with conventional methods.

Keywords: Ajax-GIS, Server-Side Rendering, Multi-layer Control, Graphical Feature Editing.

1 Introduction

GIS (Geographic Information System) [1] is a computer system to display digital map. GIS can display a variety of information on a map, and can share the information via Internet. GIS that has these characteristics is available for area marketing, disaster prevention [2], facility management, and so on.

A few years ago, a technology that is called Ajax¹ (Asynchronous JavaScript and XML) [3] appeared. Ajax technology offers a fast-loading and asynchronous updates. A typical example is GoogleMaps [4]. GoogleMaps had applied Ajax technology to web-based GIS (hereinafter referred to as "Ajax-GIS"). Ajax-GIS uses divided raster image data called "tile" in order to keep light handling. Ajax-GIS can display map by assembling tiled raster images which are transferred asynchronous from server using JavaScript program. Ajax-GIS have some issues that are attributable to using raster images. For example, it is difficult to realize a multi-layer control and graphical feature editing by using only web-browser rendering engine.

* Corresponding author.

¹ Asynchronous communication can be achieved by an XMLHttpRequest object. An XMLHttpRequest is an API that can be used by JavaScript.

To realize these functions, we propose to apply to server-side rendering for Ajax-GIS.

2 Ajax-GIS Architecture

In this chapter, we describe the basic Ajax-GIS architecture(Fig.1). Ajax-GIS comprise server side and client side. Ajax-GIS server manages several maps which are residential map, digital road map (DRM) data, disaster information, and so on. These map data are raster image data made from vector maps, and they were divided into tiled images by the rasterizing process on the server side. These raster maps are created in every scale and every layer, and managed by the server, and are used as a background map. Ajax-GIS can display the maps by means of assembling tiled raster images with the JavaScript program on the client side.

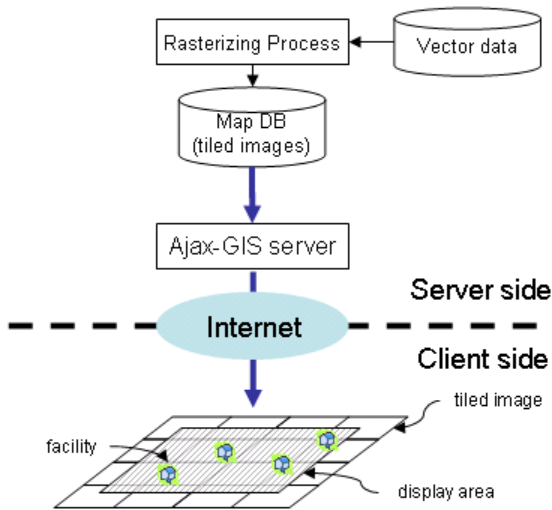


Fig. 1. Ajax-GIS operation flow

A primary characteristic of Ajax-GIS is the increased responsiveness and interactivity of web pages so that entire web pages do not have to be reloaded each time there is a need to fetch data from the server. When a user operates a map, for example scrolling the map, classic Web-GIS can't accept the user's next operation. But if Web-GIS applies Ajax technology, a user can repeat an operation without getting a reply from the server. Fig.2 shows the difference between Classic Web-GIS and Ajax-GIS. In this way, this system offers fast-loading, asynchronous display updates, and smooth map scrolling. Moreover the execution environment on the client side requires only a web-browser without particular plug-in software (e.g. Java Runtime Environment [5], Flash [6]), making this system highly convenient.

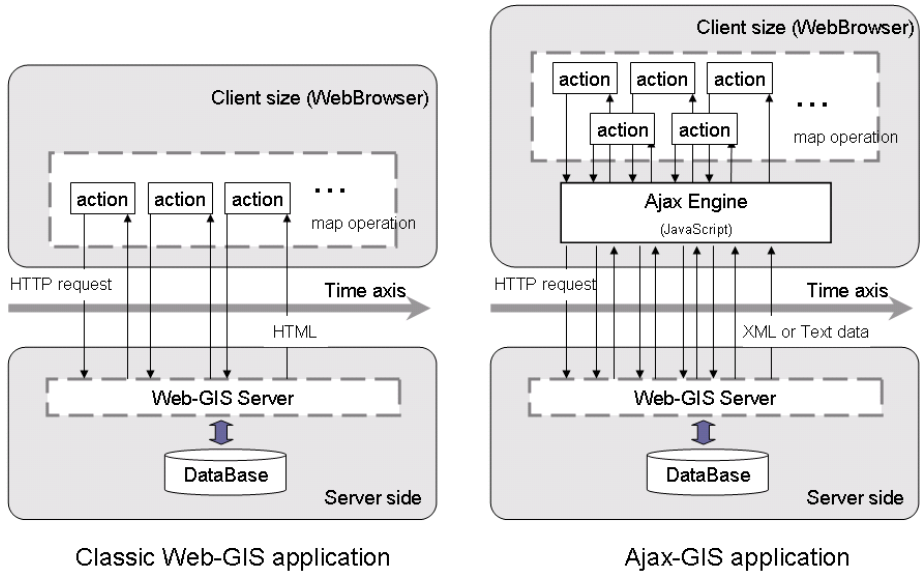


Fig. 2. Classic/Ajax Web-GIS model

3 Multi-Layer Control

In this chapter, we describe multi-layer control for Ajax-GIS. We propose that the multi-layer control is realized by means of merging tiled images in the server application as the requests of the client application on the fly.

3.1 Conventional Method

Typical implementation for Ajax-GIS realizes the multi-layer control by means of overlaying some transparent tiles at the client side (hereinafter referred to as "conventional method1"). With low number of overlaid images, this is simple and efficient. However, if larger quantities of layers are displayed by this way, the html tree that stores the image data will grow large, then the performance of map operation as a map scroll will be slow. As a result, Ajax-GIS client application should draw many tiled images. The sums of images are determined by multiplying the number of tiles by the number of layers. We developed the sample application to examine this conventional method1. As a result, we found that the display time is proportional to the number of layers. In this test, the number of layers is more than five, the map display feel very slow.

On the other hand, there is another conventional method to create image from vector data in real time by server side (hereinafter referred to as "conventional method2"). However, this method has issues that processing load becomes larger and created images are difficult to reuse by reason of client particular operations.

3.2 Proposed Method

In order to solve these issues, we apply the server-side rendering to Ajax-GIS. Fig.3 shows the architecture of multi-layer control by proposed method. At server side, tiled images with transparent background are created from map database every layer and every scale, and these images are stored in tiled image database. At client side, client map are displayed by assembled tiled images.

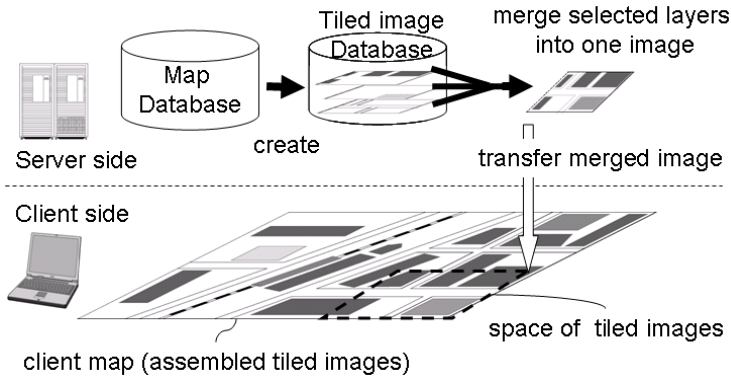


Fig. 3. Multi-layer control by proposed method

Fig.4 shows the multi-layer control process. At client side, client (1) selects display layers. This display layer information is transferred server via Internet using URL parameter. URL parameter contains display layer names. At server side, after received this information, if there is no tiled image that is selected layer in the tiled image database, the server (2) creates tiled images of selected layers. If all layers images are created in advance, this process can skip. Then server (3) merges selected tiled images into a image. Then server (4) returns the image that is overlaid selected layer images to a client. At client side, client application (5) displays merged images.

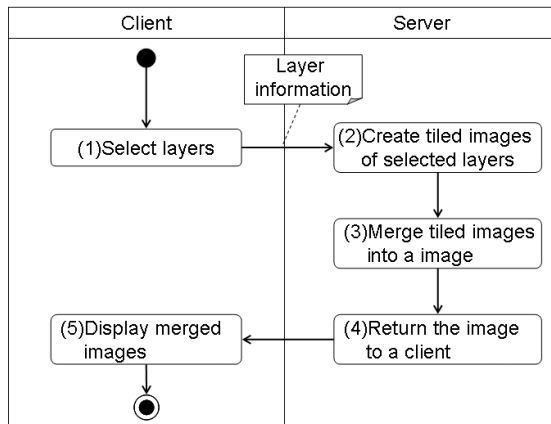


Fig. 4. Multi-layer control process

3.3 Evaluation

We developed a test application, and verified the effectiveness of our proposed method. The evaluation process is done with a JavaScript automatic map navigator. The path which is randomly generated is broken into 100 drag-and-drop steps where after each step (mouse release), queries for layer operations and for images are dispatched. As a result, we found that the proposed method is 1.5 times faster than the conventional method2 (Table.1).

Table 1. Graphical feature editing protocol

	Number of tiles	Total processing time(s)	Average processing time per one tile (ms)	Total data size of tiles (kb)
conventional method2	1160	78.9	68.0	3105.5
proposed method	1190	50.9	42.8	3101.1

4 Feature Drawing Function

In this chapter, we describe the feature drawing function for Ajax-GIS. We propose that the graphical feature editing protocol that sent from a client and send back to an image in order to edit a feature.

4.1 Conventional Method

In conventional method, graphical editing feature was realized by drawing function which is implemented in web-browser. Some rendering engines such as VML (Vector Markup Language)[7], SVG (Scalable Vector Graphics)[8] and Canvas are implemented in a web-browser. However conventional method has a main issue that implementation of drawing functions by JavaScript is dependent on these rendering engines. Therefore, we should implement drawing functions for some web-browsers. As a result, software development costs and maintenance costs will be increased.

4.2 Proposed Method

In order to solve this issue, we apply the server-side rendering. Server-side rendering can centralize drawing operations at a server, so a dependence issue associated with web-browser will be solved. This proposed method has merits which are not to need cross browser compatibility. When a client operates the feature editing such as moving coordinates or changing color, these operations are transferred as a URL parameter such as GET or POST. In order to draw feature, we defined the specific protocol to communicate between client and server. Table.2 shows the some protocol elements.

Table 2. Graphical feature editing protocol

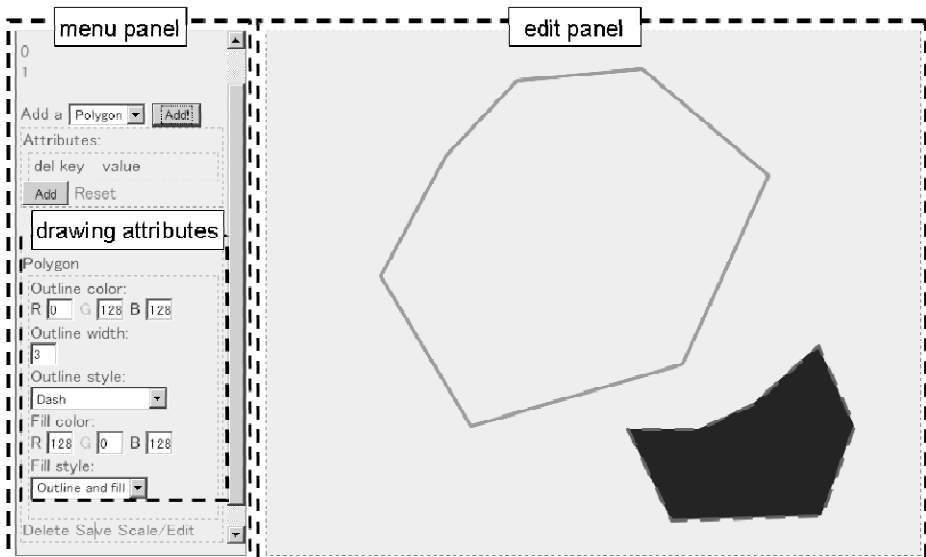
Key	Contents	Overview
cid	Client identification ID	identify uniquely the client
oid	Editing feature ID	identify uniquely the feature
t	Editing feature's coordinate ID	identify uniquely the editing feature's coordinate ID
lc	Line color	line color(e.g. blue, black)
fc	Face color	face color(e.g. gray, white)
w	Line width	line width(e.g. 2point)
ls	Line style	line style(e.g. dotted line)
fs	Face style	face style(e.g. fill)

For example, this protocol is used such as below.

```
http://localhost:80?cid=1555372038&oid=2&op=L&t=n20003&v=n20003[1067,1388]&i=1193.5,1351.5,136.5,117.5&lc=255,143,0&fc=255,255,255&w=3&ls=0&fs=0&s=0.5
```

4.3 Geometry Editor Infrastructure

We developed the prototype called "Geometry Editor" that is implemented our proposed method. Geometry Editor is a web-based editor currently supporting polyline and polygon editing through node manipulation. Nodes can be inserted, removed, and moved while an object can be shifted, scaled, and rotated in its entirety. The editor aims to support easy integration with other applications. There are two separate blocks of html code that must be included in the target application, one being

**Fig. 5.** Geometry Editor Screen

the component defining the editable region, the other being the component defining the property editor.

Fig.5 shows the Geometry Editor screen. The screen consists of two parts: a left side, displaying the menu panel that has some components to edit editing parameters (i.e. outline color, outline width, fill color, and so on), and the edit panel to draw shapes on the right side. When a client draws shapes on the edit panel, the client sends the information written by the editing protocol to the server. After the server receives the information, and then creates an image by means of server side rendering, and returns it to the client. We describe about this process in detail in the following section.

Fig.6 shows the HTML div structure of the editor. The layers related to the editor are contained in the editor layer which is in turn contained in the layer of the application to be integrated with.

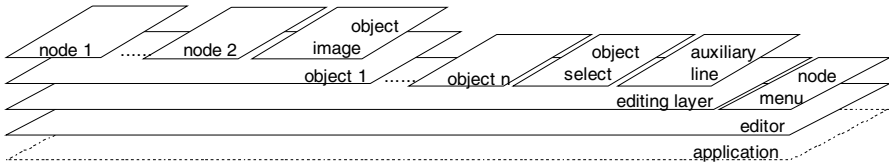


Fig. 6. Geometry Editor HTML Architecture

4.4 Ajax-GIS application

We built the Geometry Editor into our Ajax-GIS. Fig.7 shows the feature editing process between client and server. First, client (1) selects the layer of the editing target, then (2) selects region in order to narrow down the target. Next, client (3)

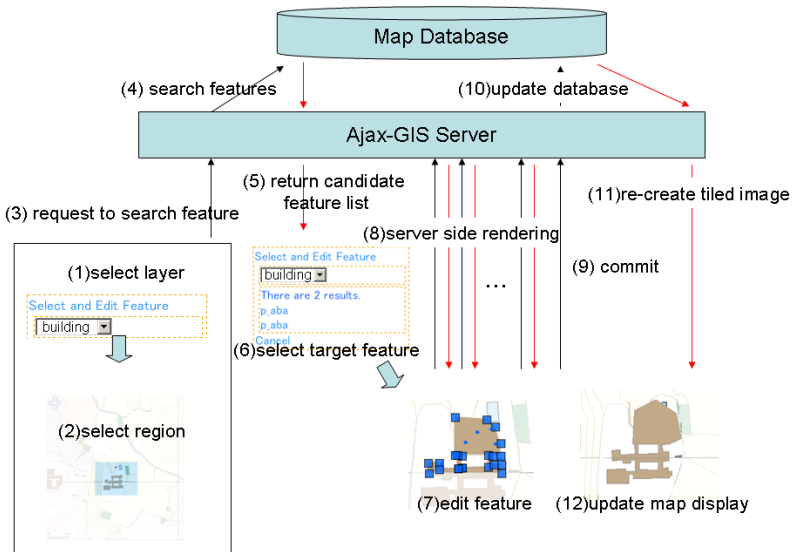


Fig. 7. Feature Editing Process

sends this selection condition to the Ajax-GIS server. After Ajax-GIS receives the selection condition, Ajax-GIS server (4) searches for features from Map database, and then (5) return candidate editing feature list to the client. The client (6) selects a target feature from the list. Then the client can (7) edit a target feature by (8) server side rendering using defined protocol. After finishing editing, the client (9) commits update data to Ajax-GIS server. When Ajax-GIS server receives the update request, Ajax-GIS server (10) updates the Map database. Then Ajax-GIS server (11) re-creates tiled image including this feature, and sends this tiled image to the client. Finally the client (12) updates map display.

For example, a time to generate a polygon fill hundreds of coordinates on the server side was less than 100ms. As a result, the proposed method has been real-time editing feature without user stress.

5 Conclusion and Feature Works

In this paper presents the methods of the multi-layer control and the graphical feature editing by the server side rendering in Ajax-GIS. In the future, we will seek further efficiency and rationalization of our method, and enrich the graphical feature editing protocol. And we will work toward practical application of a real system.

References

1. Aronoff, S.: Geographic information systems: a management perspective, p. 58 (1989)
2. Sakairi, T., Tamada, T., Nakata, H.: GIS Crisis-management System using Ajax Technology. In: SICE Annual Conference 2008, August 20-22, pp. 3043–3046. The University Electro-Communications, Japan (2008)
3. Crane, D., Pascarello, E., James, D.: Ajax In Action. Manning Publications (2005)
4. GoogleMaps, <http://maps.google.co.jp>
5. Java, <http://www.java.com>
6. Flash, <http://www.adobe.com/software/flash/about/>
7. VML, <http://msdn.microsoft.com/en-us/library/bb263898.aspx>
8. SVG, <http://www.w3.org/Graphics/SVG/>

A Method for Discussing Musical Expression between Music Ensemble Players Using a Web-Based System

Takehiko Sakamoto, Shin Takahashi, and Jiro Tanaka

University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki, Japan
sakamotot@iplab.cs.tsukuba.ac.jp,
{shin,jiro}@cs.tsukuba.ac.jp

Abstract. Music ensemble players discuss musical expression of the piece of music they perform, and determine how to play each note in a score such as the length and the dynamics of tone or phrases in every detail of the music. This paper introduces our system that supports the discussion about musical expressions on the web. Our system enables the users to write comments, draw symbols, and link videos on the score where they are discussing about. We also conducted an informal usability study to evaluate the usefulness of the system.

1 Introduction

Most music ensembles, groups of musicians, discuss musical expression to make their performance more impressive. They consider every detail of the piece of music they perform, i.e. the length, the dynamics, the articulation such as accent or staccato, and so on. They also discuss together what they feel and imagine in the piece of music and how the music is written by the composer. However, it is time-consuming to discuss them face-to-face. When they meet together, they prefer taking time to practice playing together rather than discussing. Therefore, there is a great demand for a system that enables the discussion about musical expressions asynchronously and remotely.

Current BBS and other web-based systems for remote and asynchronous discussions have problems to use for discussing musical expressions. First, since they are basically text-based system, it is difficult to specify the part in the score where the user wants to discuss. Specifying the position in the score in text requires lengthy explanation. Second, it is difficult to express various musical notations in text, since they are usually represented graphically. Third, it is not easy to refer to sound and video data.

This paper proposes a method to describe musical expressions and describes our web-based asynchronous discussion system. In our system, the users proceed discussions based on writing or drawing directly on the scores, so the users can always associate all discussions with the scores. The users can post text-based messages, musical symbol, freehand drawings and links with images and videos hosting sites. Our system can be also used for the purpose of sharing decided

directions of music ensemble i.e. fingering, timing of breath, and so on. Our system has three features:

Distant and Asynchronous Web-based System. Our system provide distant and asynchronous communication especially for amateur ensembles that cannot spent enough time for practicing. We implement our system as a web application so that the user can use various devices such as PCs and Tablets.

Discussion Along the Context of a Musical Score. In our system, the users write comments on a musical score, which makes them easy to specify the part they want to discuss about.

Visualized Description of Articulation. Articulation in the field of music means adjusting the shapes of tones, phrases and links of tones by put dynamics, emotion, and so on. General musical symbols (i.e. staccato, tenuto, portato, etc.) only represent basic articulation, not precise one in detail. In our system, the user can draw articulation “shapes” in various colors to exprese their feelings.

2 Related Work

2.1 Asynchronous Collaboration

In the studies of asynchronous discussion using graphical annotations on web browser, Heer et al. developed *sense.us* that supports asynchronous collaborative information visualization[1]. *sense.us* provides information visualization of census data in the U.S. and communication system for analyzing on web browser. Phalip et al. also developed the system that supports filmmaker and composer to make videos at a distance[2]. The feature of the system is web-based video editing sequencer that can attach annotations to its screen. Farooq et al. developed the system for scientific and creative discussion between distant places[3]. The system, named *BRIDGE* that means Basic Resources for Integrated Distributed Groupe Environments, provide virtual spaces containing timeline, chat rooms and representation to share the users’ opinions. Cadiz et al. researched communication between web editors by implementing the system that can write notes directly to web pages and making an long term experiment for a few hundred participants[4]. Ellis and Groth developed an asynchronous communication system that can attach annotations including texts, images and sounds to the video[5].

We supported distant and asynchronous discussion concerned with musical expression. Our system possesses faculty to attach annotation along the context of a musical score and to represent concrete shapes as image of the music.

2.2 Computer-Supported Ensemble

As the study of computer-supported musical performance, Bellini et al. researched supporting of real-time music performance[6]. They developed *MOODS*:

a cooperative editor for musical scores that can automatically synchronize written notes on their sheets of music. Sawchuk et al. developed *DIP* that enables players to participate their practice at a distance[7]. Akoumianakis et al. developed the prototype toolkit for the purpose of distant and asynchronous ensemble practicing[8]. The system *DIAMOUSES* records performance every part of the music and enable the users to enjoy distant ensemble by using the recorded data.

This study does not support realtime performance but discussions of the performance. We focused on discussions that is a significant part of ensemble practice, and had an eye to indirect supporting musical performance.

3 Web Client Interface for Discussing Musical Expressions

This system enables ensemble players to discuss musical expression anytime anywhere not only when they are together. In this system, the users can post their comments and reply to the other users' comments on a musical score. They can also describe graphical tone shape as their images of articulation and represent pictures or videos for their argument. The users can write musical symbols to a musical score directly in order to explain or write the result of the discussion.

Fig.1 shows the interface of the web client in this system. Tool palette(Fig.1 [A]) and help bar(Fig.1 [B]) are fixed at the left and the top of the browser. Clicking the tool palette changes the mode of this interface. There are three modes: "view", to view comments of others and to reply to them, "post", to create new comment regions on a musical score, and "stamp", to write musical symbols on a musical score. The help bar always displays advices and tips depending on the mode of the system. On a musical score(Fig.1 [C]), there are comment regions(Fig.1 [D]) and comments(Fig.1 [E]). Clicking a comment region shows or hides the list of comments there.

Comments on a Musical Score. In our system, all comments are attached to a specific region on a musical score. This makes it easy to specify the part in the musical score the comment is referring to without writing a lot of words.

When the user clicks the "post" in the tool palette, the user can create a new comment region to post a new comment. A comment region is created by a dragging operation on the musical score. When the user drags on the musical score, the system displays a comment region represented as a red and transparent rectangle (the top of Fig.2). After the dragging operation, the system also displays a comment posting box where the user can enter a text message and an articulation figure (the bottom of Fig.2). The comment posting box consists of an area for a text message, a checkbox for whether articulation is posted, and three buttons labeled "articulation", "post", and "cancel". The articulation button starts the editing an articulation figure. The user can post the entered text comment to the server by clicking the post button, or cancel it by clicking the cancel button.

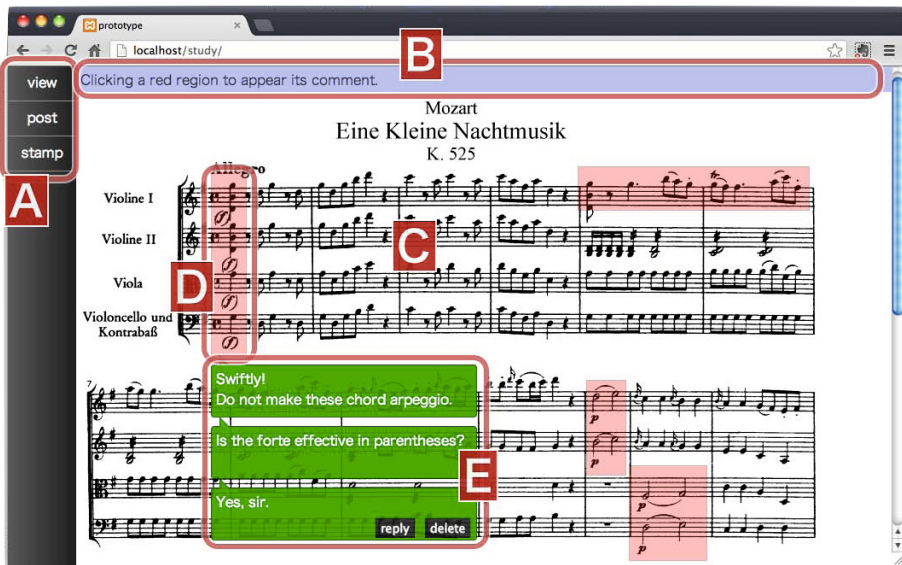


Fig. 1. Web Client Interface. [A]Tool palette. [B]Help bar. [C]Musical Score. [D]Comment region. [E]List of comment belonging to the region([D]).

A posted comment is displayed as shown in Fig.3. It has two buttons labeled “reply” and “delete”. By clicking the reply button, a new comment posting box appears, where the user can enter a reply comment. The delete button deletes the comment. The comments to the same region are listed vertically as shown in Fig.1[E].

Articulation Figures. To represent and post the image of how to play the specific part in a musical score, our system provides an interface to draw an articulation figure. The articulation figure represents the image of the volume variation and the color of a specific phrase. The user can easily edit and color it only with dragging operations.

Clicking the articulation button in the comment posting box starts the editing of an articulation figure (Fig.4). The figure in the articulation editing box represents the dynamic transition of the phrase. The left end of the figure corresponds to the attack of the phrase and the right end corresponds to the release of it.

The attack and release expression of the figure changes continuously by dragging up and down on or near each end. The color of the articulation figure can be also changed continuously by dragging at the middle. The brightness is changed by dragging vertically and the hue is changed by dragging horizontally.

After the editing of the figure, the user can post it with the comment by checking the checkbox at the left of the articulation button(Fig.4). The articulation figure is resized to fit in the comment box (Fig.3).

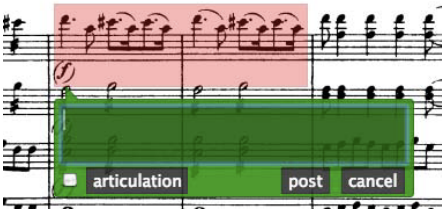


Fig. 2. Comment region(higher, red) and comment posting box(lower, green)

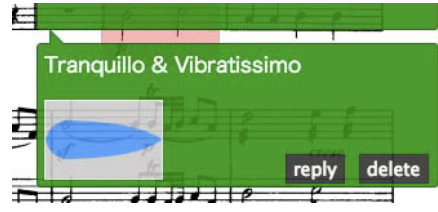


Fig. 3. Posted comment including edited articulation

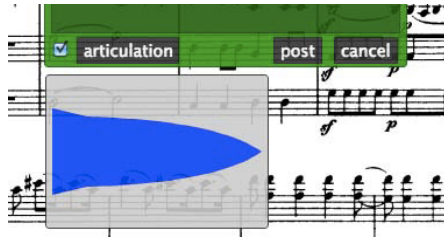


Fig. 4. Articulation editing box under the comment posting box

Referring to Images and Videos on the Web. In discussing musical expressions, it is very important to know the background of the piece of music deeply such as the character of the composer, especially in classical music. So, often the users want to refer to the other web site pages in their comments. In addition, ensemble players often think of the other players' performance as examples of how to play a music. Referring to videos is also a necessary function for our discussion system.

Therefore, in our system, the user can link to images and videos on the web to interchange these knowledge. If the URL of an image or a video is included in the entered comment, it is automatically displayed in the comment box (Fig.5). Currently, URLs of PNG/JPEG/GIF files and YouTube videos are supported in our system.

Stamping Musical Symbols. Ensemble players often write various musical symbols on music scores in order to remember how to play the part in the performance. Musical symbols are better to recognize than a textual annotation when playing a music.

Our system also provides an interface to put musical symbols on a musical score. The user can open the stamp palette (Fig.6) by clicking "stamp" in the tool palette. Then the user chooses one in the stamp palette, and put it on the musical score just like stamping.

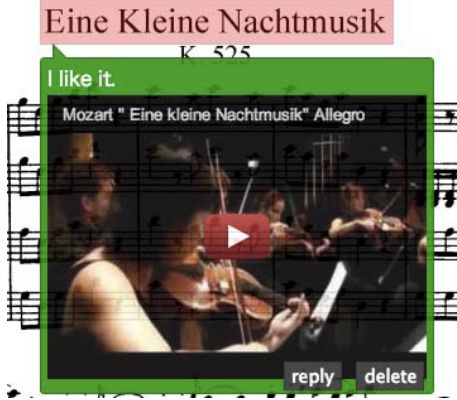


Fig. 5. Representing a video in comment

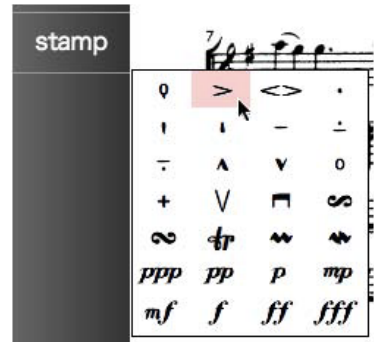


Fig. 6. Stamp palette beside tool palette

4 Implementation

4.1 System Overview

This system consists of the server and the web client(Fig.7). The web client is implemented as a web application in HTML, CSS and JavaScript for client-side, PHP for server-side and MySQL for operating database. The system benefits from HTML5 canvas and jQuery. Comment regions and comment are implemented by unordered list of HTML, and click events are controlled by JavaScript / jQuery. The database stores the informations of comment regions, comment and stamps. When the user starts this application or posts new comment, the application loads or saves them.

4.2 Implementation of the Web Client Interface

Comment on a Musical Score. When the user clicks the “post” in the tool palette, canvas element that covers the musical score will be active by changing the CSS. Dragging on the canvas, a red and transparent rectangle is dynamically drawn according to the coordinates of the mouse events in jQuery. When the new comment region is created, the system appends a new HTML li element that size and position are same as the red and transparent rectangle.

Every comment and comment posting box are formed like a balloon by CSS. When the user clicks a comment region, the list of comment belonging to the comment region toggles between expanded and collapsed states. All information of the comment and comment region are stored in database such as position, size, text message, articulation, timestamp, etc.

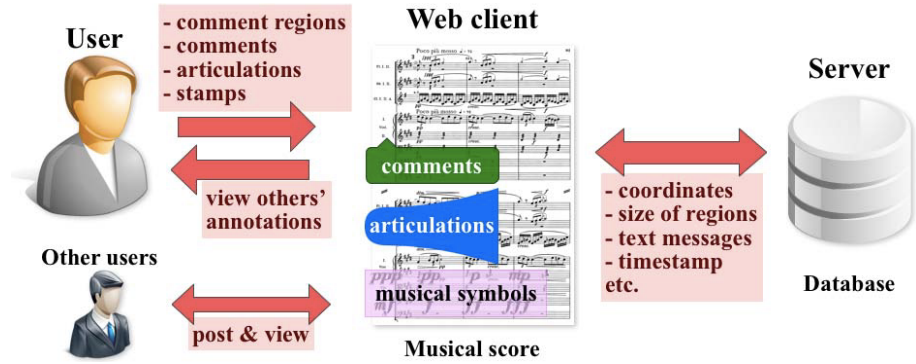


Fig. 7. System overview. The users post their opinion to musical scores. Every annotation is attached to specific region on musical scores. The database maintains data of all annotations.

Representing Image of Articulation. Editing articulation is implemented as HTML canvas. The methods of `lineTo`(straight line), `quadraticCurveTo`(quadratic Bézier curve) and `fill`(paint the closed area) render the shape of articulation. The system has several parameters that decide attack, release, brightness and hue of the shape. These parameters continuously increases or decreases according to the movement of dragging on the canvas(Fig.8). The system uses `toDataURL` method of HTML canvas to convert the image of articulation into PNG image which is displayed in a comment.

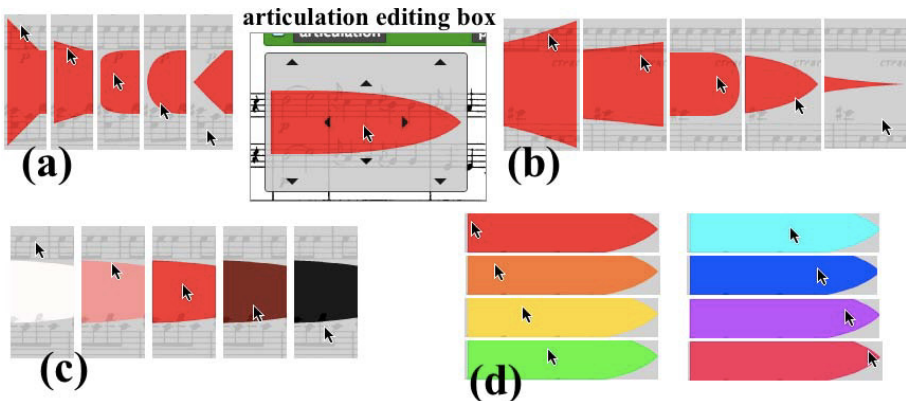


Fig. 8. Editing the articulation by dragging operation. (a)Reshaping its attack by dragging up & down at left end. (b)Reshaping its release by dragging up & down at right end. (c)Changing its brightness by dragging up & down in the middle of the shape. (d)Changing its hue by dragging left & right in the middle of the shape.

Representing Images and Videos. When the user posts a new comment, the system checks the value of textarea for containing image or video data. If the value of the textarea contains .png, .jp(e)g or .gif, the system convert the string into HTML img element. If the value of the textarea also contains URL of YouTube, the system convert the string into HTML iframe element that enables the users to watch the video immediately.

Stamping Musical Symbols. When the users clicks “stamp” in the tool palette, the stamp palette(Fig.6) appears, and the user clicks the musical symbol in the stamp palette, canvas element that covers the musical score will be active by changing the CSS. The selected musical symbol in the stamp palette is rendered near the mouse pointer, and the user can stamp it by clicking on the musical score using drawImage method. Then the stamped information as position, the kind of symbol, timestamp is saved to or loaded from the database.

5 Preliminary Experiment

We conducted preliminary experiment to evaluate the interface of this system and our approaches.

5.1 Procedure

The subjects were six university students including one female and five male. All of them has played in ensemble at least three years. We used “Eine Kleine Nachtmusik” composed by W.A.Mozert downloaded from IMSLP¹ for the musical score of this system.

The task for the subjects was to use the all functions of the discussion interface: posting a new comment, replying to a comment, editing and posting an articulation figure, putting musical symbols on a score, and posting a video of YouTube.

After finishing the task, the subjects were asked to answer to the seven point Likert scale questionnaire shown below.

- (1) Usefulness of posting comments on the musical score.
- (2) Ease of posting a new comment.
- (3) Ease of replying to a comment.
- (4) Usefulness of articulation figures.
- (5) Ease of editing the shape of articulation figures.
- (6) Ease of editing the color of articulation figures.
- (7) Usefulness of ‘stamp’.
- (8) Ease of putting a stamp on a score.
- (9) Usefulness of posting videos and images.
- (10) Demand to use this system.

Finally, we asked them open-ended question about the whole system.

¹ IMSLP: International Music Score Library Project(<http://imslp.org/>)

5.2 The Result and Discussions

Fig.9 shows the result of questionnaire. The following are the answers to the open-ended question.

- I want to edit the articulation figure more in detail. For example, I want to change the length of the tone, too.
- I don't understand clearly how to change the color of the articulation figure as desired. Choosing from several colors may be better.
- It's interesting to be able to hear the sound adjusted by the articulation figure.
- Stamping is very useful for inform the decisions.
- Colored stamps is useful for discussion.
- I want to write chord names on musical scores for music understanding.
- The screen is cluttered when many comments are posted.
- I want to use this system very much when my ensemble does not has enough time to practice.

Question \ Subject	i	ii	iii	iv	v	vi	Average
(1)Usefulness of Comment.	7	5	7	7	6	7	6.5
(2)Operation of posting new comment.	7	7	6	6	7	6	6.5
(3)Operation of replying to comment.	6	7	7	6	5	7	6.33
(4)Usefulness of articulation figures.	6	7	6	6	6	5	6.0
(5)Editing the shape of articulation.	7	5	6	7	6	7	6.33
(6)Editing the color of articulation.	6	4	3	5	4	6	4.67
(7)Usefulness of "stamp".	5	7	6	6	6	6	6.0
(8)Operation of stamping.	7	6	6	6	5	6	6.0
(9)Usefulness of videos and images.	7	7	5	7	6	7	6.5
(10)Demand to use this system.	6	6	7	7	5	7	6.33

Fig. 9. Result of questionnaire

Overall, the subjects wanted to use our system (from the the question 10). The result shows that our approach that posting comments on the specific point of the musical score is welcomed by the subjects. All subjects understood how to post or reply to a comment in a short time. However, one of the subject posted a new comment to the title of the musical score, and another subject posted a new comment to the margin of the musical score. In some cases, the user may want to post a comment unrelated to the specific part of the score.

The subjects also favored the idea of articulation figures. One subject said that it is intuitive and interesting to view the continuously changing articulation figure. However, the editing interface must be improved. It seems it is difficult to change the shape and color as they wishes.

Stamping was also mostly liked by the subjects. They said stamping is useful for informing decisions to others, but it may be not so useful for discussions. In the experiment, a lot of symbols were stamped which were not related to discussions. This result tells us, if anything, this system can be used for other process regarding music ensemble beyond discussions.

From the high score for the usefulness of videos, most ensemble players feel that it is essential to listen to other ensembles' performances. Since some video postings were not related to specific part of the score, the video posting interface can be improved to enable the posting that relates to the whole score.

Some subjects add the words that they want to use this system only when they have few time to practice, but not they have plentiful time. We believe it makes this system more demanded to enhance the features using particularity of computers such as articulation figures or referring to images and videos.

6 Conclusion and Future Work

This paper proposed a method for discussing musical expression remotely. We developed a web-based discussion system that enable the user to discuss musical expressions on the musical score. We also informally evaluated the usability of the system. The result confirms the usefulness of the system and also indicated some challenges. As future work, we will conduct the user study in actual situation and improve our system based on the result of this study.

References

1. Heer, J., Viegas, F.B., Wattenberg, M.: Voyagers and voyeurs: supporting asynchronous collaborative information visualization. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2007), pp. 1029–1038. ACM (2007)
2. Phalip, J., Edmonds, E.A., Jean, D.: Supporting remote creative collaboration in film scoring. In: Proceedings of the Seventh ACM Conference on Creativity and Cognition (C&C 2009), pp. 211–220. ACM (2009)
3. Farooq, U., Carroll, J.M., Ganoë, C.H.: Supporting creativity in distributed scientific communities. In: Proceedings of the 2005 International ACM SIGGROUP Conference on Supporting Group Work (GROUP 2005), pp. 217–226. ACM (2005)
4. Cadiz, J.J., Gupta, A., Grudin, J.: Using Web annotations for asynchronous collaboration around documents. In: Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work (CSCW 2000), pp. 309–318. ACM (2000)
5. Ellis, S.E., Groth, D.P.: A collaborative annotation system for data visualization. In: Proceedings of the Working Conference on Advanced Visual Interfaces (AVI 2004), pp. 411–414. ACM (2004)
6. Bellini, P., Nesi, P., Spinu, M.B.: Cooperative visual manipulation of music notation. ACM Transactions on Computer-Human Interaction (TOCHI) 9(3), 194–237 (2002)
7. Sawchuk, A.A., Chew, E., Zimmermann, R., Papadopoulos, C., Kyriakakis, C.: From remote media immersion to Distributed Immersive Performance. In: Proceedings of the 2003 ACM SIGMM Workshop on Experiential Telepresence (ETP 2003), pp. 110–120. ACM (2003)
8. Akoumianakis, D., Vellis, G., Milolidakis, I., Kotsalis, D., Alexandraki, C.: Distributed collective practices in collaborative music performance. In: Proceedings of the 3rd International Conference on Digital Interactive Media in Entertainment and Arts (DIMEA 2008), pp. 368–375. ACM (2008)

A Study on Document Retrieval System Based on Visualization to Manage OCR Documents

Kazuki Tamura, Tomohiro Yoshikawa, and Takeshi Furuhashi

Nagoya University, Furo-cho, Chikusa-ku, Nagoya, 464-8603, Japan

Abstract. Recently, the digitization of paper-based documents is rapidly advanced through the spread of scanners. However, tagging or sorting a huge amount of scanned documents one by one is difficult in terms of time and effort. Therefore, the system which extracts features from texts in the documents automatically, which is available by OCR, and classifies/retrieves documents will be useful. LDA, one of the most popular Topic Models, is known as a method to extract the features of each document and the relationships between documents. However, it is reported that the performance of LDA declines along with poor OCR recognition. This paper assumes the case of applying LDA to Japanese OCR documents and studies the method to improve the performance of topic inference. This paper defines the reliability of the recognized words using N-gram and proposes the weighting LDA method based on the reliability. Adequacy of the reliability of the recognized words is confirmed through the preliminary experiment detecting false recognized words, and then the experiment to classify practical OCR documents are carried out. The experimental results show the improvement of the classification performance by the proposed method comparing with the conventional methods.

1 Introduction

Scanners and printers containing function as a scanner are growing popular, so that the digitization of paper-based documents to handle them on computers is widely advanced today. Since Japanese companies were permitted to save documents electronically which were obligated to be preserved by law in 2005, a lot of paper-based documents have been rapidly digitized. And in general consumers, a large amount of document data has been stored to read them electronically because of the spread of low-cost scanners and tablet computers. Moreover, the opportunity is growing to manage various types of documents in a lump due to the diffusion of Cloud Computing. On the other hand, the larger the number of accumulated documents becomes, the more time and effort to find the required document(s) by a user are needed.

Most electronic documents on a computer have various attributes not just texts, so that document retrieval can be available by using that information. However, the scanned documents do not have even text information because they are treated as images. Texts can be embedded into digitized documents by

using Optical Character Recognition (OCR), which is the software to import printed characters into computers. OCR text usually contains some recognition errors and conversion errors. Although the recognition accuracy of OCR has been improved in recent years, that is still low in blurred documents that are often printed in the old days or handwritten ones[1]. In particular, OCR errors are more likely to occur in languages which have large number of types of characters such as Japanese and Chinese[2]. However, it is almost impossible to correct these enormous OCR errors manually. Likewise, tagging or sorting a huge amount of scanned documents one by one is also difficult in terms of time and effort. Therefore, this study aims to construct the system to retrieve the required documents by a user based on their text information from a huge amount of ones which contain texts with OCR error and are tagged without their attribute information.

Probabilistic Latent Semantic Analysis (pLSA)[3] and Latent Dirichlet Allocation (LDA)[4], which are called “Topic Model,” are the method to extract the feature of documents. Topic Model is known as the method to catch latent “topics” in the documents with high performance, which are inferred by the number of appearance of words in each document. Topic Model has been mostly applied to the documents containing correct texts so far. However, it is reported that the performance for the topic inference by Topic Model declines under the application to the documents containing OCR errors[5]. This paper studies the application of LDA to Japanese OCR documents and proposes the method to improve the performance of topic inference.

This paper focuses on that the parts incorrectly recognized by OCR are appeared as unnatural alignments in the language. Those parts are expected to become low probability in the N-gram probability obtained from large corpus. This paper defines the reliability of the recognized words, and it proposes the method weighting the appearances of words in LDA based on their reliability.

This paper supposes the system to grasp similarities among documents visually. The system calculates the distance between documents after the inference of topic distributions and visualizes them onto two-dimensional space. Two experiments are carried out by using actual OCR documents embedded classification label. First, adequacy of the reliability of the recognized words is confirmed as the preliminary experiment by the detection of false recognition. Then the classification of the OCR documents is carried out and it evaluates the classification accuracy how much the documents with correct label are located closely. The proposed LDA weighted by the reliability is compared with the conventional one, and the result shows the significance of the proposed method.

2 Related Work

A lot of studies on the information retrieval using Topic Model have been reported[6][7][8]. However, most of them suppose the application to the documents containing correct texts, not noisy texts. Although some studies use OCR texts in the experiments, most of previous work ignore the presence of

OCR errors or only attempt to remove low frequency words[9][10]. Walker et al. studied the effect of the presence of OCR errors on the performance of Topic Model. This research applies LDA to OCR documents with varying the recognition rates. The result shows that the more OCR errors appear, the more the performance of topic inference descends. However, the specific method to solve this problem is not mentioned. In addition, no study assuming the application of Topic Model to Japanese OCR documents in spite of that many studies apply it to Japanese documents recently[11][12].

On another front, the term weighting method in LDA is proposed by Wilson et al.[13]. However, the aim of this research is to improve the performance by diminishing the effect of high frequency words and function words, which is distinct from the aim of this paper.

3 Latent Dirichlet Allocation

The LDA algorithm models the D documents in a corpus as the mixtures of latent T topics where each topic generates the distribution of V words. The original LDA proposed by Blei et al.[4] assumes the appearance of topics as multinomial distributions and introduces Dirichlet distributions to their prior distribution. The LDA expanded by Griffiths et al.[14] introduces the Dirichlet distributions into the distributions of words. This paper employs the Griffiths's LDA. The generative process of LDA is represented as follows:

1. For each topic $t \in \{1, \dots, T\}$,
 - (a) Draw word distribution, $\phi_t \sim \text{Dir}(\beta)$
2. For each document $i \in \{1, \dots, D\}$,
 - (a) Draw topic distribution, $\theta_i \sim \text{Dir}(\alpha)$
 - (b) For each word $j \in \{1, \dots, N_i\}$,
 - i. Draw topic, $z_{i,j} \sim \text{Mult}(\theta_i)$
 - ii. Draw word, $w_{i,j} \sim \text{Mult}(\phi_{z_{i,j}})$

In the above process, $\text{Dir}(\cdot)$ and $\text{Mult}(\cdot)$ means the Dirichlet distribution and the multinomial distribution, respectively. α and β are the Dirichlet hyperparameters, and N_i is the number of tokens in document i .

Topics are inferred via collapsed Gibbs sampling[14]. The algorithm calculates by repeatedly sampling z_l , which is the topic of l_{th} token, based on the distribution conditioned on the values of all other elements in \mathbf{z} except z_l . N_{ijt} denotes the number of tokens of word j assigned to topic t in document i , and the summation over all values of an index is indicated with (\cdot) . And the number of tokens excluded l_{th} is represented as $N_{i(\cdot)t}^{-l}$. The sampling formula of topic inference is:

$$p(z_l | z_{\setminus l}, \mathbf{w}) \propto \frac{N_{(\cdot)jt}^{-l} + \beta}{N_{(\cdot)(\cdot)t}^{-l} + V\beta} \cdot \frac{N_{i(\cdot)t}^{-l} + \alpha}{N_{i(\cdot)(\cdot)}^{-l} + T\alpha} \quad (1)$$

After a sufficient number of iterations for updating topics, the MAP estimators of the topic distributions for all documents $\boldsymbol{\theta}$ and the word distributions for all

コミュニテイ	noun, common noun, *, *	コミュ	noun, common noun, *, *
システム	noun, common noun, *, *	=	noun, verbal noun, *, *
		デイステム	noun, common noun, *, *
(a) Result of “R jeBVXe”		(b) Result of “R =fCVXe”	

Fig. 1. Example of Morphological Analysis by MeCab

topics ϕ can be obtained. When θ_t^i and ϕ_j^t represent the generated probabilities of topic t in document i and word j in topic t , these are calculated by eq. (2).

$$\theta_t^i = \frac{N_{i(\cdot)t} + \alpha}{N_{i(\cdot)(\cdot)} + T\alpha} \quad \phi_j^t = \frac{N_{(\cdot)jt} + \beta}{N_{(\cdot)(\cdot)t} + V\beta} \tag{2}$$

4 Proposed Method

4.1 Purpose

Topic Model estimates topics in documents using the information of the words and their occurrences. The languages such as Japanese or Chinese which do not have space between words must be applied morphological analysis in order to divide them into words. However, inadequate divisions frequently occur in noisy texts, then the words acquired by the morphological analysis are also inadequate.

Here we describe an example that OCR confused “R jeBVXe” with “R =fCVXe” in Fig. 1. The results were obtained by MeCab¹, the most popular Japanese morphological analysis engine, are shown in Fig. 1. In Fig. 1(b), the parts of false recognition are separated as noun inadequately. This paper connects the successive nouns and unknown words and treats as a combined word. Then, the reliability on the recognition of the word is defined based on the probability of collocation using N-gram probability, and the reliability is introduced to topic inference. First, we define the reliability of the recognized words. Then, we propose the weighting method of topic inference based on the reliability.

4.2 Reliability of Recognized Words

The probability of word-level N-gram is that a certain N words (morphemes) occur contiguously. The N-gram probability is obtained from a large corpus, and it can be said that high probability pattern is general and natural sequence of words but low probability pattern is unnatural. This paper employs word-level Bi-gram ($N=2$) to calculate the reliability of words. In the case that a word w consists of the morphemes $t_1t_2 \cdots t_n$, the Bi-gram probability of the word w is:

$$p(w) = p(t_1) \prod_{i=2}^n p(t_i|t_{i-1}) \tag{3}$$

¹ <http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html>

The reliability of word w_i is described as $m(w_i)$, which is defined by the geometric mean in combination probability in w_i $p_{\bar{t}}(w_i) = p(w_i)^{\frac{1}{n}}$ as eq. (4). W denotes the all words in the documents. When $p(w)$ is 0, $m(w)$ is assigned to 0.

$$m(w) = \frac{\log p_{\bar{t}}(w_i) - \arg \min_{w \in W} \log p_{\bar{t}}(w)}{\arg \max_{w \in W} \log p_{\bar{t}}(w) - \arg \min_{w \in W} \log p_{\bar{t}}(w)} \quad (4)$$

4.3 Term Weighting LDA

Wilson et al. developed term weighting LDA (TW-LDA)[13], which weights on the words in LDA and infers topics. It can be considered that the weighting in this method corresponds to the expansion of the multinomial distribution into real number, while it is not specified in [13]. In the conventional LDA, the word and the topic of l_{th} token are shown as V and T dimensional vector described “1-of- K .” “1-of- K ” is the vector in which only one element is “1” and the others are “0.” TW-LDA allocates the real number instead of “1” into the element to reflect weights of words in the topic estimation. M_{ijt} denotes the sum of the weights of tokens of word j assigned to topic t in document i . Topics are inferred via collapsed Gibbs Sampling, and the formula is:

$$p(z_l | z_{\setminus l}, \mathbf{w}) \propto \frac{M_{(\cdot)jt}^{-l} + \beta}{M_{(\cdot)(\cdot)t}^{-l} + V\beta} \cdot \frac{M_{i(\cdot)t}^{-l} + \alpha}{M_{i(\cdot)(\cdot)}^{-l} + T\alpha} \quad (5)$$

This paper uses the reliability of the recognized words described in 4.2 as the weight of TW-LDA. It is expected that the emphases of high reliability words make the performance of topic inference improved in poor OCR documents.

5 Experiments

5.1 Applied Documents

The experiments employed some documents picked out from the proceedings of the 74th National Convention of Information Processing Society of Japan. Two data sets were picked out: Data 1 was 31 documents and they consisted of four sessions, and Data 2 was 48 documents and they consisted of six sessions. Each data set had the texts without errors embedded electronic documents correctly and with OCR errors obtained from the scanned documents. Some noise level were prepared as the OCR texts by adding various degree of noise onto the documents artificially. The session belonging to was regarded as the correct classification label. We assume that users usually manage both originally digitalized texts and OCR texts. Thus the documents with OCR errors and without errors were mixed 50:50 in this experiment. Error rate of OCR was measured by PER (Position-independent Word Error Rate)[15], which is a common metric of the performance in the areas such as machine translation. The PER of the OCR documents without artificial noise was about 0.25.

Table 1. Comparison of performance to extract false recognized words

(a) Data 1							(b) Data 2						
PER	Low frequency			Reliability			PER	Low frequency			Reliability		
	P	R	F	P	R	F		P	R	F	P	R	F
0.28	0.305	0.924	0.459	0.568	0.674	0.616	0.23	0.292	0.910	0.442	0.569	0.763	0.652
0.57	0.452	0.956	0.613	0.756	0.809	0.782	0.34	0.443	0.952	0.604	0.728	0.765	0.746
0.53	0.508	0.936	0.659	0.806	0.754	0.779	0.43	0.530	0.954	0.682	0.782	0.803	0.792
0.60	0.508	0.958	0.664	0.837	0.794	0.815	0.51	0.601	0.948	0.736	0.823	0.770	0.795
0.65	0.568	0.947	0.710	0.879	0.804	0.840	0.57	0.626	0.969	0.761	0.858	0.813	0.835
0.70	0.566	0.957	0.711	0.823	0.807	0.815	0.66	0.670	0.937	0.781	0.896	0.808	0.850
							0.70	0.692	0.940	0.797	0.917	0.814	0.863

In these experiments, Adobe Acrobat² was used as the OCR software and morphological analysis was carried out by using MeCab. N-gram probability was derived from Web Japanese N-gram Version 1³.

5.2 Preliminary Experiment

This preliminary experiment was done to investigate the adequacy of the reliability of words defined in 4.2. The experimental dataset described in 5.1 were employed. We picked out 1000 words in total randomly from each noise level and labeled them whether correct recognition or false one manually. Two methods were compared: one was the removal based on the reliability of words in 4.2 and the other was that of low frequency words[9]. The former extracted the words with the reliability below the threshold⁴ and the latter extracted the words appeared once in the documents. Precision and recall of the false words in the extracted words were calculated, and F-measure, the harmonic mean of the precision and the recall, was acquired. The results of the comparison are shown in Table 1. In Table 1, precision (P), recall (R), and F-measure (F) are listed.

In the point of precision, the method removing low frequency words was much worse than that based on the reliability, because the removing method also removed a lot of words recognized correctly. By contrast, the reliability method could extract the false words appropriately because this method is independent of the frequency. On the other hand, recall of the reliability method became lower than the removing method. The reason was that the most of false words were occurred only once, so they could be removed by the frequency. In Table 1, the holistic evaluation index F-measure shows that the proposed reliability method performed much better than the removing method. Therefore, it was confirmed that the low reliability words in the proposed reliability of words represented false words adequately.

² Adobe Acrobat 9.46, <http://www.adobe.com/jp/>

³ Taku Kudo, Hideto Kazawa, "Web Japanese N-gram Version 1", published by Gengo Shigen Kyokai

⁴ The threshold was set to 0.30 by the result of the prior experiment.

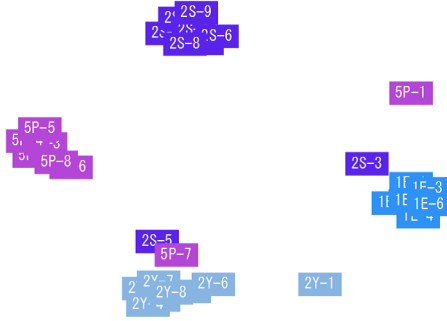


Fig. 2. Example of visualization of Data 1. “2Y-1” represents the proceeding number and the color of a node represents the corresponding session.

5.3 Experiment of Documents Classification

The comparison of the proposed method with the conventional methods was carried out based on the accuracy of documents classification shown in 5.1.

Visualization System. Here we explain the visualization system which was used in this experiment. First, this system estimates the topic distribution of each document. Then, distances between documents are calculated based on the topic distributions. This system employed Jensen-Shannon divergence[16] which is the symmetric index of Kullback-Leibler divergence and is known as the distance between probabilistic distributions as the index of distance[17].

$$D(d_i, d_j) = \sum_{k=1}^K \left[p(z_k|d_i) \log\left(\frac{2p(z_k|d_i)}{p(z_k|d_i)+p(z_k|d_j)}\right) + p(z_k|d_j) \log\left(\frac{2p(z_k|d_j)}{p(z_k|d_i)+p(z_k|d_j)}\right) \right] \quad (6)$$

Finally, documents are allocated into two-dimensional space. This system used Multi-dimensional Scaling (MDS)[18] as the method for the dimension reduction. MDS can preserve the distances among data in the original space as much as possible while it embeds the coordinates of data onto lower (two) dimensional space. When $l_{i,j}$ denotes the original distance between d_i and d_j and $l_{i,j}^*$ denotes the distance in the visualization space, the coordinate of each document $\chi = \{\mathbf{x}_i \in \mathbf{R}^2, i = 1, 2, \dots, D\}$ is available from the minimization of the following error function:

$$E(\chi) = \sum_{i=1}^{N-1} \sum_{j=i+1}^N (l_{i,j}^* - l_{i,j})^2 \quad (7)$$

In this function, $l_{i,j}^*$ is given by the Euclidean distance, $l_{i,j}^* = \|\mathbf{x}_i - \mathbf{x}_j\|$.

The example of visualization of Data 1 is shown in Fig. 2.

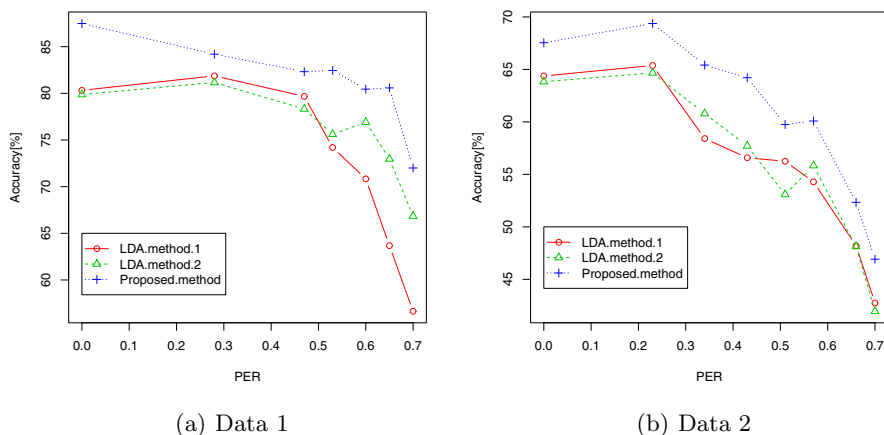


Fig. 3. Comparison of classification accuracy in each PER. Horizontal axis is PER, and Vertical axis is the classification accuracy.

Evaluation Measurement. The purpose of this visualization system is to support a user to understand documents' relationship intuitively. Thus, we evaluated the result of visualization based on the accuracy of the labels predicted with the k -nearest neighbor (k -NN) method in the visualization space[8]. This method predicts a label of a document by a majority vote of correct labels of k nearest samples. Classification accuracy was defined as the ratio of the number of labels predicted correctly to all documents. This accuracy becomes higher when the same label documents are allocated in the neighborhood and different ones are allocated far each other. This experiment was applied with $k = 5$. The accuracy of Fig. 2 was 84%.

Experimental Condition. The hyperparameters of the Dirichlet distributions in LDA and TW-LDA were $\alpha = 0.1$ and $\beta = 0.1$, and the sampling was run for 1000 iterations. The classification accuracy of 50 trials was averaged by the reason the topic inference depends on the initial topics given randomly. The number of topics was determined by the best one in terms of the accuracy in electronic documents in the prior experiment: $T = 4$ in Data 1 and $T = 6$ in Data 2, which were the number of sessions (correct labels) as a result. We compared with two conventional methods: one was a general LDA (called "LDA method 1") and the other was a LDA after the removal of words appeared once described in 5.2 (called "LDA method 2").

Results and Discussions. The results of classification accuracy are shown in Fig. 3. It was observed that the accuracy of LDA method 1 was declined drastically at around 0.5 in Data 1 and around 0.3 in Data 2. This result sup-

port with the description of Walker et al.[5] in which the performance of LDA descends by the presence of OCR errors. Though LDA method 2 shows little improvement around $PER = 0.6 - 0.7$ in Data 1, overall results were almost same with LDA method 1. By contrast, the proposed method worked better than the conventional LDA methods in each error rate of both data sets. There were the significant differences between LDA method 1 and the proposed method and between LDA method 2 and the proposed method, respectively, which were confirmed by the paired t-test considering multiplicity by Sidak's statistical method ($p < 0.01$). Particularly in Data 1, the proposed method could keep higher accuracy at the error rates in which the performance of LDA method 1 got drastically worse. This result shows that the proposed method preserved the depression caused by OCR errors.

Although it was confirmed that the accuracy was improved by the proposed method, accuracy tended to decline in all methods according to decreasing the word error rate. It is thought to be due to the approach of the proposed method that was not to correct error words but to diminish the effect of error words. The use of the information of error words which would be effective for LDA if they were recognized correctly. We will try the corrections of error words and the reflection of their information such as similarity in shape to the topic inference.

6 Conclusion

This paper studied the application of LDA for Japanese OCR documents to extract their features and proposed the method to reduce deterioration of the performance of topic inference caused by OCR errors. The proposed method focused on the unnatural alignment in the part of false recognition, and quantitatively defined the reliability of the recognition of words. Then it put weights on words in LDA based on the reliability and tried to improve the performance of topic inference.

First, the preliminary experiment was done to confirm the adequacy of the defined reliability of the recognized words by the detection of false recognition, and showed that the proposed reliability could extract the error words better than the conventional preprocessing which removed low frequency words. Then the significance of the proposed method was shown comparing with the conventional methods in the classification performance.

We will study the correction of error words and the use of the information to improve the performance of topic inference more. Also the extraction of significant words and putting weights on them will be discussed. In addition, we try to construct the document retrieval system based on the proposed method.

References

1. Bunke, H.: Recognition of cursive roman handwriting: past, present and future. In: Proc. Seventh International Conference on Document Analysis and Recognition, pp. 448–459. IEEE (2003)

2. Nagata, M.: Japanese ocr error correction using character shape similarity and statistical language model. In: Proc. of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, vol. 2, pp. 922–928. Association for Computational Linguistics (1998)
3. Hofmann, T.: Probabilistic latent semantic indexing. In: Proc. of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 50–57. ACM (1999)
4. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. *The Journal of Machine Learning Research* 3, 993–1022 (2003)
5. Walker, D., Lund, W., Ringger, E.: Evaluating models of latent document semantics in the presence of ocr errors. In: Proc. of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 240–250. Association for Computational Linguistics (2010)
6. Wei, X., Croft, W.: Lda-based document models for ad-hoc retrieval. In: Proc. of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 178–185. ACM (2006)
7. Yao, L., Mimno, D., McCallum, A.: Efficient methods for topic model inference on streaming document collections. In: Proc. of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 937–946. ACM (2009)
8. Iwata, T., Yamada, T., Ueda, N.: Probabilistic latent semantic visualization: topic model for visualizing documents. In: Proc. of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 363–371. ACM (2008)
9. Blei, D., Lafferty, J.: Dynamic topic models. In: Proc. of the 23rd International Conference on Machine Learning, pp. 113–120. ACM (2006)
10. Newman, D., Block, S.: Probabilistic topic decomposition of an eighteenth-century american newspaper. *Journal of the American Society for Information Science and Technology* 57(6), 753–767 (2006)
11. Yokoyama, S., Eguchi, K., Ohkawa, T.: Distilling information diffusion networks from blogosphere using latent topics. *The IEICE Transactions on Information and Systems (Japanese Edition)* 93(3), 180–188 (2010)
12. Kitajima, R., Kobayashi, I.: An examination for proper events grasping latent topics in a document and its application. *IPSJ SIG Notes NL-201(3)*, 1–8 (2011)
13. Wilson, A., Chew, P.: Term weighting schemes for latent dirichlet allocation. In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, HLT*, vol. 10, pp. 465–473 (2010)
14. Griffiths, T., Steyvers, M.: Finding scientific topics, vol. 101, pp. 5228–5235. *National Acad. Sciences* (2004)
15. Och, F., Tillmann, C., Ney, H., et al.: Improved alignment models for statistical machine translation. In: Proc. of the Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora., pp. 20–28 (1999)
16. Lin, J.: Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory* 37(1), 145–151 (1991)
17. Heinrich, G.: Parameter estimation for text analysis (2005)
18. Kruskal, J.: Nonmetric multidimensional scaling: a numerical method. *Psychometrika* 29(2), 115–129 (1964)

Audio-Visual Documentation Method for Digital Storytelling for a Multimedia Art Project

Chui Yin Wong, Chee Weng Khong, Kimberly Chu,
Muhammad Asyraf Mhd Pauzi, and Man Leong Wong

Universal Usability and Interaction Design (UID) SIG,
Interface Design Department, Faculty of Creative Multimedia,
Multimedia University, 63100 Cyberjaya, Malaysia
{cywong, cwkhong, kimberly.chu, asyraf.pauzi, mlwong}@mmu.edu.my

Abstract. In this paper, we describe an interactive multimedia art project, namely *FaceGrid*, using mosaic photography art concept for digital storytelling. Inspired by mosaic photography and a montage concept, *FaceGrid* was produced by using many small image tiles that were woven and stitched together to form the pixel art design pattern. *FaceGrid* documents the different ways of living and lifestyles of ordinary folks in a multi-cultural and diverse ethnic society in Malaysia. We use audio-visual documentation methods (photography and film-documentary techniques) to record, capture and archive the different facets of lives and user stories by ordinary people. We then transform those slices of life via digital storytelling technique into an interactive multimedia art project.

Keywords: Audio-visual documentation method, digital storytelling, multimedia art, ordinary folks.

1 Introduction

In the past decade, the prevalence of various new media platforms has enabled ordinary people like us being able to be seen or heard on the ‘news’. On many occasions, people like to share, to hear, to read, to check and to receive everyday news. We also like to share stories with others about ourselves, our experiences, and about the news we hear and read through various forms of media such as newspaper, radio, television, blogs, and social networks (i.e. Facebook, Twitter). As a result, the exclusivity and media coverage received by celebrities and politicians can now be shared by ordinary folks who, in this project, are provided an opportunity to share their snippets of their life stories and experiences via convergent media.

2 Motivation: Why Ordinary Folks

Ordinary communities living in a developing nation such as Malaysia have invaluable life stories, culture and experiences. A multi-cultural and multi-ethnic society have much to share about what it means to be free, to have dreams, to uphold principles,

values and religious beliefs, to have families and lives, and many of them untold. [1] described how researchers can encounter the ‘unnoticed’ as:

‘Everyday life is something we tend to take for granted, something that just is, something unnoticed. But everyday life is perhaps the most important dimension of society – it’s where we live most parts of our lives with each other.... Looking at everyday activities and experiences, from language and emotions to popular culture and leisure, encountering the everyday explores what social structures, orders and processes mean to us on a daily basis.’

As a result, *FaceGrid*¹ multimedia art project addresses the absence of the voices of ordinary folks. It aims to document and portray the extraordinary stories of the untold life stories by ordinary people in Malaysia in digital storytelling format. Based on a series of short interviews using audio-visual documentation methods (photography and videography) on Malaysians from all walks of life, the multimedia art project captures the many ways that they experience living, working, communicating, and ageing. Through the use of vignettes, a glimpse of the livelihood, cultural practices, and freedom of the population. This multimedia art project, using research-creation process, critically and creatively explores the potential of multimedia forms and formats to depict the extraordinary diversity within the so-called ordinary lives of Malaysians.

3 Research Creation Process

This research-creation project provides its audience an insight towards the intersection of storytelling, multimedia art, and interactivity. Rather than presenting single unified images of a local ordinary folk’s profile, the many dimensions, and contradictions, is presented to the audience of what it means to live in a multi-cultural and multi-ethnic community in the context of multimedia art installation and online web (www.facegrid.info) platforms.

There are 3 stages involved in the research-creation process. Firstly, we formulate the conceptual design of *FaceGrid* based on an inspiration of faces montage. Secondly, we use audio-visual documentation method to conduct a series of in-depth interviews with the local folks. Lastly, the process involves transcription, translation, video-and-photo-editing in the media production process to transform into an interactive multimedia platform.

3.1 Conceptual Design of *FaceGrid* User Interface

FaceGrid project is an interactive multimedia art project produced by an interdisciplinary research team consisting of photographers, videographers, interface and interaction designers, an ethnographer, and a sound artist. *FaceGrid* was inspired by mosaic art and is a montage of image tiles created from photography that is woven and stitched

¹ *FaceGrid* multimedia art project can be accessed online via www.facegrid.info

together to form a grid (design pattern). The image montage was then analysed and arranged, so as to match the overall concept design of its intended artistic form. With the transition of images in a pixelated grid manner, *FaceGrid* interface enables scalability and zoom-ability for the audience to interact with any multimedia content and elements (i.e. audio, images, text, video) through each art pixel. The users will experience ordinary folks' lifestyles through digital storytelling while interacting with our interactive multimedia platform, thus stirring every audience for their insights towards the diverse ethnic unity and multi-cultural integration.

There are two platforms for *FaceGrid*, which are multimedia art installation and online website version. Users are allowed to toggle, interact, drag and move the profile content "pixels" around *FaceGrid* multimedia art project. For *FaceGrid* multimedia art installation platform, an audio controller is also presented to allow the user to turn on or mute the background audio as desired (Fig. 1). Whenever each pixel is selected, users are able to view, interact, interpret and re-create with the selected pixel's (profile) content (Fig. 2). However, one will notice that every pixel of multimedia content that is shown will not be deemed as a discrete whole, but contains several irreconcilable and contradictory meanings - an amalgamation of photography, audio and video into a dense interactive format. It portrays that any one pixel should not be interpreted literally, but consists of alternative perceptions, and will invoke the audience into making various opinions and having a mixture of experiences.

This project provides its audience the different facets of ordinary folks' life stories and an insight through the intersection of digital storytelling, multimedia art, and interactivity. Rather than presenting single unified images of a local subject, the many dimensions, and contradictions, of what it means to live in a multi-cultural and multi-ethnic community is presented to the audience in context of a country that achieved its independence over 50 years ago.

Fig. 2 is a content page, which shows a Malay warrior descendent, an amateur weapon collector, whom likes to collect all the Malay weapons i.e. lady dagger, *keris*



Fig. 1. User interface of *FaceGrid* multimedia art installation project (All rights reserved)

(Malay warrior dagger) since its ancient time. In the video interview, he explained how the different type of Malay weapons were used during the ancient war period, and how those collections have been past to him from his ancestor till now.



Fig. 2. Each pixel contains different user stories in *FaceGrid* content page (All rights reserved)

The user interface for *FaceGrid* website online version (www.facegrid.info) is slightly different as compared to multimedia art installation platform. Fig. 3 allows users to key-in text for profiles searching, tagging and also allow users to leave comments. In addition, it also incorporates with social media features such as Facebook, Google Plus, Tumblr, Tweeter for profile sharing, which allows for higher publicity as compared to the multimedia art installation.



Fig. 3. A website online version of *FaceGrid* (www.facegrid.info) (All rights reserved)

3.2 Audio-Visual Documentation Method in *FaceGrid*

We document the different ways of living, experiences and work using audio-video documentation methods (photography and videography) of ordinary local folks [1, 2, 3, 4]. We seek consensus from the local folks using either oral or written consent before the interview. Subsequently, all of these inputs were transferred, re-packaged and transformed into an interactive design platform.

We use structured interview given in a set of interview questions to document the ordinary' life story; their passion towards what they are currently doing, any particular hobby/artifact/interest they have, experience of technology, and to explore their perception about being a Malaysian. Thus far, we had conducted 66 interviews (include photo-documentary). As this project is still on-going, the total user profiles will be growing as time goes along. In terms of sampling selection, we use a snowballing method to recruit our user profiles. We attempt to encompass different ethnicity as stated in Malaysian society section, with various educational and socio-economy backgrounds. Hence, we approached our relatives, colleagues, and friends whom possess different backgrounds, various culture and also diverse ethnicity. Some are also strangers whom we approached them about the notion of project through our fieldwork.

Due to diverse cultural society of Malaysia, we usually conduct the interview in various languages. For those who can converse English fluently as spoken language, the interview was conducted in English whereas would be conducted in their mother tongues such as Mandarin, Cantonese (a dialect of Southern China, normally spoken by Hong Kong citizens), and Malay language. We then convert those local languages into English sub-title in our video captions during our multimedia production phase.

During the research-creation process, we employ photo-documentation [4] as one of the visual research methods in visual methodology. Unlike most of the visual methods where the researcher studies images created by artists e.g. photo-elicitation method [4], photo-documentation method invites the researchers make the images with their own creation in different forms such as film, video, photographs, maps, diagrams, paintings, drawings, collages. In photo-documentation, a researcher takes a carefully planned series of photographs to document and analyze a particular visual phenomenon [4, p. 298]. Rose [4] mentioned that photo-documentation method is not commonly employed by researchers as compared to photo-elicitation method, which makes this project unique and more authentic as our own research creation.

In our context, we extend the photo-documentation method, and employ audio-visual documentation method to document their ways of living by using a combined photography and videography techniques. Each profile was informed about their disclosure of identity when briefed the consent form and purpose of project. With their consensus, we took photograph of their hobbies that they like to share with others (i.e. artifacts, material objects), their living or work environment, family and beloved photos (See an example of Fig. 4). For some photos which we could not capture on the spot, we asked the interviewee whether s/he can share their old photos, which they considered as memorable in their lives. We then scan and return those photos to them after the transcription and video-editing work.

Images such as photographs are seen as valuable for the local folks profile in portraying their individual life story. Collier [1] claimed that ‘photographs are precise records of material reality’, and photography are taken in a systematic way in order to provide data which the researcher then analyses and interprets its meaning. Besides, [1] also argued the use of photographs alongside interviews. Grady [5] also stated that ‘pictures are valuable because they encode an enormous amount of information in a single representation’. According to Becker [6], photos are valuable for the way they convey ‘real, flesh and blood life’, especially making their audiences ‘bear witness’ to that life [7].

In the *FaceGrid* project, multimedia art installation and online website platforms share the similar content elements e.g. photos, videos and text description except user interface and its arrangement. The following describes the user interface of *FaceGrid* multimedia art platform, in which a profile content interface is divided into three regions (Fig. 4). Region A: Photo Collage portrays some snippets of life stories of the user profile. Region B is a brief text description to provide users or audience the profile’s background information. Region C allows users to toggle the interview in a video form. For instance, the photo collage at the left panel in Fig. 4 tells the relation of cultural artifacts for a 75 year-old trishaw peddler (Uncle Ah Peng) at the UNESCO world heritage site, Malacca historical town. During the interview, the informant shared his passion of spending over RM10,000 for ornament decorations on his trishaw as his hobby with the motive for welcoming his passengers as hospitality of Malaccan tourism. He was so proud when he talked about his trishaw, and how he decorated his trishaw given thoughts on his ornaments. For example, the top right photo in the photo collage below (see Region A as in Fig. 4) shows a Malay yellow ‘*tengkolok*’ (headdress), which represents the symbolism of Malay Sultanate

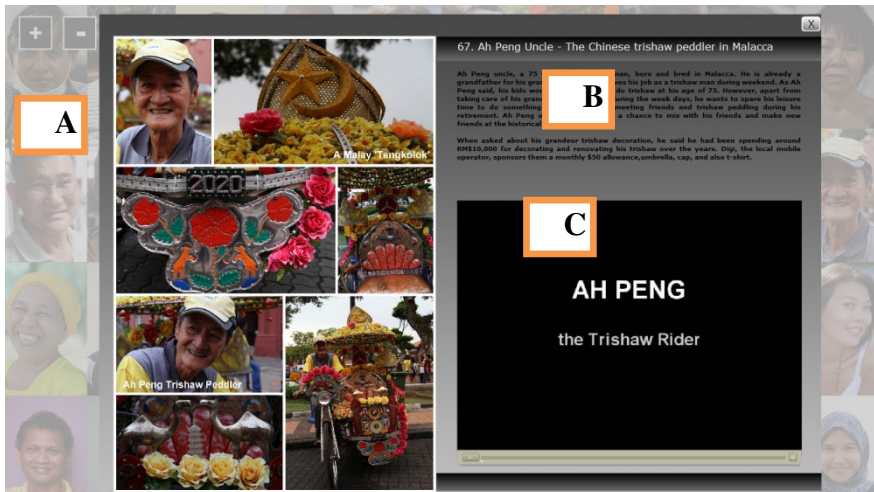


Fig. 4. A profile content page that portray a snippet of a senior’ ordinary life stories in *FaceGrid* multimedia art installation platform (All rights reserved). A: Photo Collage. B: Brief Text Description. C: Video Interview (All rights reserved).

(Emperor)'s royalty. The mouse deer and hibiscus flower (in second row left in photo collage) symbolizes the State of Malacca and national flower respectively. The camels (below left in the photo collage) denote the mobility of a vehicle in ancient time that carries passenger, which indirectly signifies the meaning of trishaw fetching the passengers from one place to another place. Region B of the right top panel denotes a brief description of the informant.

It is arguable that the criterion of images must be considered carefully taken, and the audience who view photographs and interpret the meaning of the photographs is always context-specific. We agreed that the role of photograph itself has to be clearly present, either as self-evident evidence or as evidence whose significance is established through the research-creation process [4]. More importantly, in photo-documentation method, photographs are used to examine the social effects of images with which this method is most concerned centered on the relations between the researcher, those people they are researching, and the photos [4].

Apart from the visual photography, the session is usually followed by an in-depth interview of each local folk's profile that is documented using videography. (This is shown on Region C: Video Interview in Fig. 4.) The purpose is for the users to have a view of understanding the local folks sharing their life story and views of livings and work in documentary film. The interview is structured interview with several sections, which are: the local folk's life story, perception on current livings, work, hobbies, interest or artifacts, challenges in life, and perception being a Malaysian. Before end of the interview session, they are also given a chance for them to share their life wisdom or experience that gives advice to others.

According to Appadurai [8], "we follow the things themselves, for their meanings are inscribed in their forms, their uses, their trajectories. It is only through analysis of these trajectories that we can interpret the human transactions and calculations that enliven things. Thus, even though from a theoretical point of view human actors encode things with significance, from a methodological point of view, it is the things-in-motion that illuminate their human and social context."

From the kaleidoscopic intersection of many voices, a collective tale emerges from a combination of distinct narrative fragments [9]. Narrative is all around us when we share our experiences, our beliefs, our history, our wellbeing, etc. in various means be it in static or moving images, audio, video or in print [10]. In this project, we transform the user profile's life stories into a digital storytelling platform in the *FaceGrid* multimedia art using audio-visual documentation method.

4 Users Interaction with Facegrid

For *FaceGrid* multimedia art installation, a user can use a touch pad to interact with the individual user profile. *FaceGrid* will be projected out on a wide screen (using white cloth or portable projector screen) with projector at the background, and audio-visual multimedia system (Fig. 5). The grid somewhat acts as a structured manner in an artistic form to provide the users a vehicle to delve, and to participate in a stranger's experiences on a computer-based platform.



Fig. 5. Users/Audience interacting with *FaceGrid* multimedia art installation using a touchpad

5 Conclusion

This paper demonstrates the many facets of lifestyle, work and living in a multi-ethnic nation of Malaysia in an interactive multimedia art platform known as *FaceGrid*. As the country celebrates its 55th independence anniversary, this work commemorates the freedom, the livelihood, the struggle, the biodiversity, and the developments experienced by Malaysians in their everyday lives. This project is an on-going exercise accumulating the community sharing of sustenance, experiences, careers, history, artefacts and many more on a unified platform as a means for sharing and disseminating information. It uses a multimedia approach in capturing and documenting, and illustrates these captured moments in an interactive platform known as *FaceGrid*.

Certainly, we have extended the photo-documentation method [4] to encompass videography into audio-visual documentation method in this research-creation project. As stated by Rose [4], photo-documentation method is not widely used in the interaction design discipline and visual research studies, so does audio-visual documentation method. As the project is still on-going, we still have rooms to argue and debate the viability of its research method in the context of interaction design and visual methodology. However, we believe audio-visual documentation method can serve as a rigorous visual method in design process, with careful way of documenting visual and aesthetical appearances and relating them to social processes, in which the researcher engages in a reflexive discussion of the coding in the research-creation process.

Acknowledgments. We are indebted to the informants, Malaysian communities and the public on the whole for sharing and for allowing us to document their life stories,

working lifestyles, history and experiences. In addition, we also like to thank the research assistants, Yiing Y'ng Ng, Aaron Pang, Ivy Yap and Hui Zhen Tan, whom assist for the media production. This project is funded by TM R&D and Canadian SSHRC research grants.

References

1. Collier, J.: *Visual Anthropology: Photography as a Research Method*. Holt, Rinehart and Winston, New York (1967)
2. Heath, C., Luff, P., Hindmarsh, J.: *Audio Visual Methods in Social Research*. Sage, London (2009)
3. Jacobsen, M.H. (ed.): *Encountering the Everyday: An Introduction to Sociologies of the Unnoticed*. Palgrave Macmillan, Basingstoke (2009)
4. Rose, G.: *Visual Methodologies: An Introduction to Researching with Visual Materials*, 3rd edn. Sage, London (2012)
5. Grady, J.: Working with visible evidence: an invitation and some practical advice. In: Knowles, C., Sweetman, J. (eds.) *Picturing the Social Landscape: Visual Methods and the Sociological Imagination*. Routledge, London (2004)
6. Becker, H.: Visual evidence: A seventh Man, the specified generalization, and the work of the reader. *Visual Studies* 17, 3–11 (2002)
7. Holliday, R.: Reflecting the self. In: Knowles, C., Sweetman, J. (eds.) *Picturing the Social Landscape: Visual Methods and the Sociological Imagination*. Routledge, London (2004)
8. Appadurai, A.: Introduction: commodities and the politics of value. In: Appadurai, A. (ed.) *The Social Life of Things: Commodities in Cultural Perspective*. Cambridge University Press, Cambridge (1986)
9. Meadows, M.S.: *The art of interactive narrative*. Pearson, USA (2002)
10. Bolter, J.D., Gromala, D.: *Windows and mirrors: Interaction design, digital art, and the myth of transparency*. MIT Press, Cambridge (2003)

Author Index

- Abdelnour-Nocera, José III-146
Abed, Mourad V-141
Abou Khaled, Omar IV-157, IV-167,
V-19
Acharya, Subrata II-26, III-3
Ackermann, Daniel III-446
Ahn, Jong-gil II-249, III-87
Akihiro, Ogino V-372
Akiyoshi, Masanori I-159, II-444
Alexandris, Christina IV-3, IV-13
Alfredson, Jens I-221
Almutairi, Badr IV-23
Altaboli, Ahamed I-549
Alves, Aline Silva III-324
Alves, Paulo Henrique Cardoso II-3
Amandi, Analía A. III-107
Amer, Hossam III-352
Amin, Rahul III-97
An, Kwang-Ok V-3
Anastasiou, Dimitra IV-32
Andujar, Marvin II-335
Angelini, Leonardo II-531, V-19
Angulo, Julio III-10
Ann, Sangmin V-441
Arantes, Ana Cláudia Costa II-370
Ariya, Kanako I-275
Armentano, Marcelo G. III-107
Arning, Katrin III-49
Arnold, Todd I-593
Arya, Ali IV-215
Asaba, Nobutake I-565
Atia, Ayman II-561
Attard, Judie V-122

Bach, Cédric I-521
Badioze Zaman, Halimah IV-523
Baguma, Rehema III-249
Bahn, Sangwoo IV-561, IV-594, IV-610,
IV-618
Baranak, Andrew III-285
Barbé, Jérôme IV-429
Barbosa, Simone Diniz Junqueira I-3,
I-460, IV-439
Barnett, Julie I-166

Barroso, João V-39
Beldad, Ardion III-371
Bell, Carrie III-285
Bellini, Pierfrancesco III-259
Belluati, Maurizio III-277
Benetti, Angelo II-428
Benford, Steve II-376
Bengler, Klaus II-578, II-596
Bergmann, Francine B. III-117
Bergmans, Anne II-541
Bernhaupt, Regina I-521
Bevan, Nigel I-281
Beyerer, Jürgen IV-252
Bilgen, Semih I-310
Billman, Dorrit I-193
Bim, Sílvia Amélia I-3
Blezinger, David V-10
Böck, Ronald V-301, V-381
Borsci, Simone I-166, I-203
Borum, Nanna III-418
Boscarioli, Clodis I-3
Bourimi, Mohamed III-39
Boy, Guy Andre IV-629
Brat, Guillaume I-290
Braun, Andreas IV-147
Braun, Roman III-446
Brejcha, Jan I-13
Bröhl, Christina III-127
Brooks, Anthony L. III-418
Brouwers, Aurélie III-136
Brunk, Sören V-46
Brunner, Simon IV-388
Bruno, Ivan III-259
Burghardt, Manuel I-176
Burzacca, Paolo I-241
Bützler, Jennifer III-127

Cagiltay, Nergiz E. I-310
Câmara, Marcela IV-439
Campos, José Creissac I-421
Canedo, Arquimedes I-350
Cantú, Andrea II-316
Caon, Maurizio II-531, IV-167, V-19

- Cardozo, Aline Teodosio dos Santos II-370
 Carrino, Francesco II-531, IV-157
 Carrino, Stefano II-531, IV-167
 Caruso, Mario IV-637
 Catarci, Tiziana IV-637
 Cederholm, Henrik V-403
 Cerrato-Pargman, Teresa II-464
 Chakraborty, Joyram II-13
 Chalon, René V-29
 Chamberlain, Paul I-22
 Chan, Susy III-183
 Chang, Hsien-Tsung III-359, IV-177
 Chang, Hsin-Chang II-606, II-654
 Chang, Tsen-Yao III-237
 Charoenpit, Saromporn II-343
 Chen, C.L. Philip V-236
 Chen, Huan-Ting IV-177
 Chen, Po Chun V-132
 Chen, Wei-Ju II-20
 Chen, Weiqin IV-186
 Chen, Xiuli I-193
 Cheng, Cheng IV-243
 Cheng, Mingkai IV-243
 Cheng, Yun-Maw II-20
 Cheung, Yiu-ming V-246
 Chien, Tsung-Tien IV-177
 Ching, Joanna Sin Sie II-181
 Chizari, Mahmoud V-262
 Cho, HyunKyoung I-32
 Cho, Yongjoo V-339
 Choh, Ikuro IV-348
 Choi, Heesun IV-618
 Christensen, Line G. III-418
 Chu, Chi Nung II-386
 Chu, Kimberly IV-750
 Chuang, Ming-Chuen I-540
 Chung, Byung Do III-402
 Chung, Ching-Wen III-199
 Chung, Donghun II-258
 Cincotti, Febo IV-637
 Cirilo, Elder José Reoli II-3
 Clamann, Michael IV-551
 C. Mac, Khoi-Nguyen IV-67
 Coats, Brian II-26
 Coleti, Thiago Adriano I-184
 Corbett, Brendan IV-561
 Craig, Claire I-22
 Craven, Michael P. II-36, II-189
 Crebolder, Jacquelyn I-109
 Crowe, John II-36, II-189
 Cruciol, Leonardo L.B.V. IV-648
 Cusseau, Guillaume II-46
 Dahlbäck, Nils II-586
 Das, Amitava V-310
 Davaasuren, Enkhbat IV-196
 David, Bertrand V-29
 de Abreu Braz, Priscilla Fonseca IV-439
 de Araújo, Fábio Rodrigo Lopes II-3
 Debattista, Jeremy V-122
 de Carvalho, Alfredo Veiga II-3
 Dehlinger, Josh II-66
 Delomier, Florent V-29
 de Lucena, Carlos Alberto Pereira II-127, II-144, II-370
 de Lucena, Carlos José Pereira II-3, II-370
 De Marsico, Maria II-351
 Dempster, Peter III-269
 Deng, Xiaoming V-227
 de Oliveira, Helvecio Siqueira II-428
 de Oliveira, João Batista S. III-117
 de Oliveira, Káthia Marçal I-211
 DePalo, Philip II-56
 de Souza, Clarisse Sieckenius I-320
 Dey, Anind K. V-92
 Dharma, Anak Agung Gede IV-568, IV-578
 Di Bitonto, Pierpaolo II-484
 Dicke, Christina II-551
 Di Giovanni, Pasquale III-342
 Dixon, Jeremy II-66
 Dixon, Shannan DeLany II-66
 Dörner, Ralf I-381
 Drury, Jill L. IV-658
 Duan, Fuqing V-206, V-216, V-280
 Duong, Anh-Duc IV-513
 Ebert, Achim I-371
 Ebisawa, Yoshinobu IV-205
 Edirisinghe, Chamari IV-601
 Ehrhart, Stefan I-371
 Eibl, Maximilian IV-706, V-196
 Ekandem, Josh I. II-335
 Engel, Jürgen I-300, I-401
 Eriksson, Jeanette V-403
 Erturan, Yusuf Nasuh I-310
 Esat, Ibrahim II-76, V-262
 Eshraghi, Mona II-76

- Eshraghi, Saba II-76, V-262
 Eskildsen, Søren II-361
 Espinoza, Matías II-299
 Etemad, S. Ali IV-215
 Euman, Rob III-146
- Fain, Brad III-285
 Fallman, Daniel I-128
 Fang, Xiaowen II-150, III-183
 Farhadi-Niaki, Farzin IV-215
 Faria, Jose V-39
 Feary, Michael I-193
 Federici, Stefano I-203
 Fekry, Mohamed II-561
 Feng, Shih-Yi IV-177
 Fernandes, Hugo V-39
 Ferreira, Juliana Jansen I-320
 Ferreira, Sérgio IV-678
 Fischer, Arthur IV-147
 Fleischmann, Albert I-330, III-456
 Forbrig, Peter I-51, I-300, I-340
 Förster, Ulrich III-446
 Frajhof, Leonardo II-127, II-370
 Franke, Thomas II-612
 Fritsch, Lothar III-68
 Froehlich, Elaine IV-668
 Fuchs, Christian IV-678
 Fuglerud, Kristin Skeide I-41
 Fukumoto, Kiyotaka IV-205
 Funke, Alexandra V-46
 Furuhashi, Takeshi IV-740
 Furukawa, Hiroshi III-156
- Gabbard, Joseph L. III-217
 Gabillon, Yoann I-211
 Gadahad, Pallavi Rao II-181
 Gailliot, Ava IV-135
 Galatas, Georgios IV-43
 Galindo, Michel I-51
 Gambäck, Björn V-310
 Gao, Dekun IV-225
 Gao, Yue V-153
 Garcia, Franco Eusébio II-229
 Germanakos, Panagiotis II-418
 Ghedira, Khaled V-141
 Ghinea, Gheorghita V-56
 Gilbert, Juan E. II-335, III-97
 Giménez, Rafael I-256
 Go, Kentaro I-77, I-119, I-137, I-500,
 I-531, IV-456
- Göbel, Matthias I-61
 Gomez, Randy IV-408
 Gonçalves, Gil IV-678
 Gong, Yang II-94
 Gopalakrishnakone, Ponnampalam
 IV-601
 Goto, Yukio IV-722
 Gotoda, Naka II-84
 Gotoh, Takuto II-216
 Gotoh, Tomomi IV-235
 Greene, Kristen K. IV-449
 Grinsell, Jon II-46
 Gris, Iván IV-97
 Grønli, Tor-Morten V-56
 Gruss, Sascha V-474
 Grüter, Barbara II-323
 Gu, Zhenyu I-555
 Guercio, Elena III-277
 Guo, Ping V-236
 Guo, Weimin IV-243
- Hachimura, Kozaburo I-611
 Hamdy, Aya II-561
 Hammer, Jan Hendrik IV-252
 Han, Chia Yung IV-291
 Handschuh, Siegfried V-122
 Hansen, Jarle V-56
 Harata, Miho V-319
 Harley, Linda III-285
 Harris, Helen II-571
 Harun, Afdallyna II-376
 Hasegawa, Koyo II-647
 Hashiyama, Tomonori I-480
 Hashizume, Ayako I-68
 Hattori, Masatsugu IV-462
 Haubner, Nadia I-381
 Hayakawa, Seiji I-77, I-119, I-137, I-500,
 I-531
 Hayami, Takehito IV-312, IV-320,
 IV-330
 Hayashi, Kazuya IV-320
 Hayashi, Yugo III-20
 He, Ning V-153, V-216
 Heikkinen, Kari II-239, V-431
 Heinroth, Tobias III-59
 Helmbrecht, Magnus II-578
 Henschen, Lawrence J. II-107, IV-688
 Herberhold, Carlotta III-381
 Herdin, Christian I-300, I-401
 Hermann, Fabian I-256, III-29, V-122

- Hermann, Sven I-350
 Hernández Rubio, Erika IV-301
 Heupel, Marcel III-39
 Hijikata, Yoshinori III-314
 Hilborn, Olle V-403
 Himmel, Simon III-49
 Hiramatsu, Yuko II-398
 Hirata, Ichiro I-361
 Hishina, Masateru I-565
 Ho, Yi-Lun III-359
 Hölscher, Christoph V-10
 Hori, Maiya V-391
 Horne, Sara II-76
 Hörold, Stephan I-85, I-391
 Howes, Andrew I-193
 Hrabal, David V-474
 Hsieh, Min-Chih III-166
 Hua, Lei II-94
 Huang, Chi-Yu III-199
 Huang, Lixiao IV-561
 Huang, Xinpeng I-480
 Huang, Yi-Pai IV-506
 Huang, Yu Ting II-386
 Hufgard, Andreas III-438
 Hui, Mingqi V-498
 Humayoun, Shah Rukh I-371
 Hung, Jian-Yung II-606, II-654
 Hung, Matthew III-285
 Huseyinov, Ilham N. II-391
 Hwang, Sheue-Ling II-606, II-654,
 III-166, III-193
 Hwang, Wonil IV-587
 Hwangbo, Hwan I-103
 Hyden, Douglas IV-135

 Ichino, Junko I-480
 Ifukube, Tohru II-398
 Iida, Yusuke III-466
 Ingold, Rolf IV-157, IV-167
 Inoue, Satoshi II-647
 Ioannou, Nikolaos IV-358
 Ishihara, Makio V-163
 Ishihara, Manabu II-474
 Ishihara, Yukio V-163
 Itakura, Naoaki IV-225
 Iteya, Satoru V-411
 Ito, Atsushi II-398
 Ito, Kyoko V-421
 Itou, Junko III-174
 Iwai, Yoshio V-270, V-391

 Iwata, Mitsuru I-480
 Iwata, Naoko I-146
 Izumi, Tomoko III-295, V-329

 Jackson, France III-97
 Jakus, Grega II-551
 Jander, Hans I-221
 Jeng, Tay Sheng V-132
 Jeon, Myounghoon IV-49
 Jeon, Wooram IV-594
 Jepsen, Henrik W. III-418
 Jerg-Bretzke, Lucia V-474
 Ji, Yong Gu I-103
 Jiang, Yonghan IV-243
 Jin, Lianwen V-254
 Jochems, Nicole III-127
 Johnson, Chris II-20
 Johnson, Michael D. II-408
 Jokela, Timo II-101
 Jokinen, Kristiina IV-32, IV-262
 Jonsson, Ing-Marie II-586
 Joo, Jaekoo V-62
 José, Rui I-421
 Ju, Hehua V-171
 Jung, Tzyy-Ping V-448

 Kaber, David B. IV-551
 Kaji, Ken V-372
 Kakusho, Koh IV-340
 Kalwar, Santosh Kumar II-239, V-431
 Kambayashi, Yasushi IV-235
 Kamei, Katsuyuki IV-722
 Kampf, Constance III-388
 Kanayama, Naofumi I-480
 Kanenishi, Kazuhide II-503
 Karunanayaka, Kasun IV-601
 Kashima, Tomoko II-444
 Katagami, Daisuke IV-340
 Kato, Toshikazu III-466, V-411
 Kawabe, Akhiro III-295
 Kawahara, Tatsuya IV-408
 Kawakami, Hiroshi III-301
 Kawakatsu, Hikaru IV-491
 Kellen, Vince III-183
 Kesdoğan, Doğan III-39
 Khaled, Omar Abou II-531, V-186
 Khong, Chee Weng IV-750
 Kida, Takahiro IV-235
 Kim, Da-Hey V-3
 Kim, Dongkeun V-339, V-441

- Kim, Dongsoo IV-587
 Kim, Eun Yi V-362
 Kim, Gerard Jounghyun II-249, III-87
 Kim, Hyo Chang I-103
 Kim, Jongbae V-3
 Kim, Jonghwa V-441
 Kim, Minyoung V-339
 Kim, Youngwon II-249
 Kinoshita, Yuichiro IV-456
 Kirakowski, Jurek IV-87
 Kirche, Elias III-492
 Klein, Andreas III-396
 Klein, Gary L. IV-658
 Kleindienst, Jan IV-59
 Kline, Keith III-285
 Knecht, Christian III-29
 Ko, Deajune V-346
 Kobayashi, Noriyuki III-307
 Kobayashi, Tomoyuki IV-378
 Koh, Yoon Jeon III-402
 Kohls, Niko II-163
 Kohn, Izumi I-146
 Kojima, Kazuaki IV-398
 Kojima, Takatsugu IV-340
 Komatsu, Tsuyoshi I-585
 Kosnik, David E. II-107
 Kouroupetroglou, Georgios IV-358
 Kraft, Patricia III-411
 Krause, Michael II-596
 Krems, Josef F. II-612
 Kristensen, Kasper III-418
 Kroes, Lies II-117
 Krömker, Heidi I-85, I-391
 Kryssanov, Victor III-20
 Ku, Pei-Ying II-606
 Kühn, Romina IV-698, V-46
 Kümmerling, Moritz I-411
 Kunc, Ladislav IV-59
 Kuno, Yuki V-72
 Kuo, Chih-Chung II-606, II-654
 Kuramochi, Toshiya III-314
 Kuribara, Takuro IV-416
 Kurosawa, Toshifumi IV-272, IV-469
 Kurosu, Masaaki I-68, I-95

 Labský, Martin IV-59
 Laine, Juha II-101
 Lalanne, Denis II-531, IV-388, V-186
 Lam, Jacob III-418
 Laterza, Maria II-484

 Le, Hoang-An IV-67
 Leal, Jeremy I-109
 Leal Ferreira, Simone Bacellar III-324
 Lee, Eui Chul V-346
 Lee, Hyunji II-258
 Lee, Jaehong IV-368
 Lee, Ja Young IV-610
 Lee, Ju-Hwan IV-49
 Lee, Julia C. IV-688
 Lee, Unseok IV-281
 Lehmann, Simon I-381
 Lekkas, Zacharias II-418
 Lemme, Diana IV-698
 Leotta, Francesco IV-637
 Lepreux, Sophie I-211
 Lewis, Richard I-193
 Li, Haidong V-82
 Li, Yan IV-568
 Li, Yueqing IV-594
 Li, Yu-Wen III-359, IV-177
 Liang, Chao V-171, V-206
 Liang, Xueyi V-82
 Lietz, Holger V-196
 Lietzenmayer, Ryan IV-135
 Lim, Brian Y. V-92
 Lim, Ji Hyoun IV-368
 Lima, Gabriel Vial Correa II-3
 Lima, Tania II-428
 Limbrecht-Ecklundt, Kerstin V-301
 Limongelli, Carla II-434
 Lin, Ming-Hui IV-506, V-483
 Lin, Qing-Wen III-193
 Lin, Ray F. III-199
 Lin, Yuan-Pin V-448
 Lin, Yu-Ting V-483
 Lindley, Craig V-403
 Liu, Cheng-Li III-475
 Liu, Hai-Lin V-82
 Liu, Rong V-206
 Liu, Shijing IV-561, IV-618
 Liu, Yijun V-498
 Liu, Yikun IV-658
 Løvborg Jensen, Kasper III-332
 Lombardi, Matteo II-434
 Long, Zhiying V-498
 Longo, Lucia III-277
 Lorite, Salvador I-231
 Lou, Jian I-555
 Lu, Ke V-153, V-216
 Lucid, Brian IV-668

- Luderschmidt, Johannes I-381
 Luo, Bin V-227

 Ma, Tao IV-291
 Ma, Wenqi IV-551
 Macek, Tomáš IV-59
 Maekawa, Yasuko II-134
 Majewski, Martin IV-147
 Majima, Yukie II-134
 Makedon, Fillia IV-43
 Maki, Atsushi V-411
 Malagardi, Ioanna IV-13
 Manssour, Isabel H. III-117
 Manthey, Robert V-196
 Marani, Alessandro II-434
 Marcus, Aaron I-13
 Marinc, Alexander IV-147
 Märтин, Christian I-300, I-340, I-401
 Martin, Jennifer L. I-166
 Martin, Jim III-97
 Martinie, Célia I-51, I-290
 Martins, Paulo V-39
 Massey, Adam II-36, II-189
 Matsui, Tatsunori IV-398
 Matsumoto, Shimpei II-444
 Matsuura, Kenji II-84
 Matsuyama, Takashi IV-408
 Mattar, Nikita V-102
 Mayas, Cindy I-85, I-391
 Mbatha, Blessing II-454
 Mbatha, Mbali II-454
 McDougall, Siné I-575
 Mecella, Massimo IV-637
 Medjkoune, Sofiane IV-77
 Meixner, Gerrit I-411
 Mele, Maria Laura I-203
 Mello, Erick II-197
 Mendori, Takahiko II-521
 Meneses Viveros, Amilcar IV-301
 Mennecke, Brian III-428
 Mhd Pauzi, Muhammad Asyraf IV-750
 Micheals, Ross J. IV-449
 Miike, Katsuaki I-565
 Miles, Robert II-36
 Min, Wensheng IV-540
 Minker, Wolfgang III-59
 Miranda, Pedro Augusto da Silva e Souza II-3
 Mita, Yuusaku IV-416
 Mito, Kazuyuki IV-225

 Mitsuishi, Takashi I-565
 Miwa, Shotaro V-178
 Miyaji, Chikara II-84
 Miyaji, Yutaka I-620
 Mizuno, Tota IV-225
 Modesto, Débora Maurmo III-324
 Mohagheghi, Amir II-76
 Mollard, Régis IV-429
 Mondin, Fabio Luciano III-277
 Mont'Alvão, Claudia Renata II-127, II-144, II-370
 Monteiro, Ingrid IV-439
 Morandini, Marcelo I-184, I-249
 Moreau, Guillaume IV-530
 Moreira, Samuel I-421
 Moriwaka, Makoto IV-320
 Morreale, Patricia II-335
 Morrissey, Kellie IV-87
 Motta, Gustavo H.M.B. II-197
 Mouchere, Harold IV-77
 Mourlas, Constantinos II-418
 Mugellini, Elena II-531, IV-157, IV-167, V-19, V-186
 Munemori, Jun III-174
 Muñoz, Adolfo I-231
 Murakami, Satoru I-565
 Murata, Atsuo IV-312, IV-320, IV-330
 Murata, Kazuyoshi IV-462, IV-491

 Nagano, Hiromi V-319
 Nakagawa, Koji II-84
 Nakamura, Tetsuaki IV-126
 Nakamura, Yutaka III-156
 Nakatani, Yoshio III-295, V-329
 Nakatsu, Ryohei IV-601
 Nam, Chang S. IV-561, IV-594, IV-610, IV-618
 Neris, Vânia Paula de Almeida II-229, III-227
 Nesi, Paolo III-259
 Nguyen, Vinh-Tiep IV-67
 Nguyen-Huynh, Duy-Hung IV-513
 Nicholas, Michael II-640
 Nieminen, Marko II-101
 Nishida, Shogo III-314, IV-378, V-421
 Nishino, Kazunori II-511
 Nishiuchi, Yusuke II-521
 Noguchi, Anna IV-469
 Noor, Nor Laila Md. II-376
 Norcio, Anthony F. II-173

- Nothdurft, Florian III-59
 Nouri, Elnaz II-266
 Nouri, Jalal II-464
 Novick, David IV-97
 Novotny, Philipp V-454
 Nunes, Fátima de Lourdes dos Santos
 I-184

 Oami, Takuma IV-568
 Obata, Akihiko III-307
 Obata, Yuhki IV-568
 Oblaender, Vera IV-706
 Ochi, Keita IV-330
 Oe, Tatsuhiro IV-469, IV-712
 Ogawa, Hitoshi III-20
 Oh, Hyung Jun I-103
 Ohiro, Tomoya V-329
 Ohkawa, Yuichi I-565
 Ohkura, Michiko I-585, II-343, II-647
 Ohmori, Kosuke V-421
 Ohori, Kotaro III-307
 Okada, Akira V-353
 Okada, Naoki III-314
 Okawa, Jun V-372
 Oki, Maho III-209
 Okuuchi, Keita IV-340
 O'Malley, Claire II-376
 Ono, Yuichi II-474
 Orfgen, Marius I-411
 Otmame, Samir IV-530
 Owen, Charles B. V-464
 Özcebe, Esra I-310
 Ozturk, Elif II-408

 Palanque, Philippe I-51, I-290
 Pang, Natalie II-181
 Paolucci, Michela III-259
 Pappu, Aasish IV-107
 Paredes, Hugo V-39
 Park, Jae Heon III-402
 Park, Jongsoo III-217
 Park, Kyung Shin V-339
 Park, Kyung Eun II-56
 Park, Sangin V-346, V-441
 Parthasarathy, Rangarajan II-150
 Pastor, Enric I-231
 Paternò, Fabio I-241
 Paulick, Peyton II-163, V-454
 Peeters, Jeroen I-128
 Peng, Xiaobo II-408

 Pengnate, Supavich I-593
 Perez Pelaez, Mariano IV-348
 Peters, Anicia III-428
 Petersson Brooks, Eva III-418
 Petitrenaud, Simon IV-77
 Pfaff, Mark S. IV-658
 Pham, Truong-An IV-67
 Philippow, Ilka I-510
 Pichelmann, Stefan II-612
 Pierantoni, Felipe II-127, II-144
 Pino, Alexandros IV-358
 Platt, Donald IV-629
 Plischke, Herbert II-163, V-454
 Ponsa, Pere I-231
 Porras, Jari II-239, V-431
 Potamianos, Gerasimos IV-43
 Prates, Raquel Oliveira I-3
 Price, Chandler III-285

 Rabe, Felix IV-117
 Rabie, Osama II-173
 Rahmanivahid, Pooyan II-76, V-262
 Raine-Fenning, Nicholas II-36, II-189
 Rampoldi-Hnilo, Lynn II-309
 Ramsbrock, Jens II-621
 Randall, Tania I-109
 Raposo, Alberto IV-439
 Rauch, Marta II-276
 Rauff, Stefanie III-438
 Ray, Jerry III-285
 Rebenitsch, Lisa V-464
 Rehm, Matthias I-431, II-361, III-332
 Rekimoto, Jun V-112
 Reppa, Irene I-575
 Riccio, Angela IV-637
 Ridi, Antonio IV-157
 Rigas, Dimitrios IV-23
 Rintel, Sean III-484
 Ritter, Marc V-196
 Rivera, Ismael V-122
 Robichez de Carvalho, Gustavo II-3
 Rocha, Mario Sandro II-428
 Rodeiro-Iglesias, Javier I-441, I-490
 Rodil, Kasper III-332
 Rodrigues, Kamila III-227
 Røssvoll, Till Halbach III-68
 Romano, Marco III-342
 Roselli, Teresa II-484
 Rossano, Veronica II-484
 Rudnicky, Alex IV-107

- Ruffieux, Simon V-186
 Rukavina, Stefanie V-474
 Ryu, Taebeum IV-368

 Saito, Yukou II-647
 Sakairi, Takeo IV-722
 Sakamoto, Kiyomi V-353
 Sakamoto, Maki IV-126
 Sakamoto, Takehiko IV-730
 Sakamoto, Yuichiro IV-469
 Sakashita, Seiji V-353
 Sakata, Nobuchika IV-378
 Sakoda, Masayuki II-134
 Sakurai, Yoshihisa II-84
 Samaras, George II-418
 Sampanes, Anthony Chad II-284
 Sampanthar, Kes II-292
 Sánchez, Jaime II-299
 Sandnes, Frode Eika II-20
 Santos, Oscar Francisco dos I-249
 Santos, Thebano Almeida II-428
 Sarathy, Rathindra I-593
 Sasaki, Shiori II-493
 Sato, Ayaka V-112
 Sato, Kosuke V-270
 Satoh, Hironobu II-521
 Scerri, Simon III-29, V-122
 Scheruhn, Hans-Jürgen III-446
 Schettini, Francesca IV-637
 Schieck, Ava Fatah gen. V-10
 Schierz, Christoph II-163
 Schlegel, Thomas IV-698, V-46
 Schlick, Christopher M. III-127
 Schmidt, Michael IV-479
 Schmidt, Werner I-330, III-456
 Schnädelbach, Holger II-36
 Schneidermeier, Tim I-176
 Scholz, Sebastian C. I-451
 Schuller, Andreas I-256, III-29, V-122
 Schwaller, Matthias IV-388
 Schwarz, Markus J. V-454
 Sciarrone, Filippo II-434
 Sebillo, Monica III-342
 Segers, Mariël III-371
 Segura, Vinícius I-460
 Seki, Hironori II-647
 Seki, Makito V-178
 Selvarajah, Kirusapillai II-36, II-189
 Seo, Ki-Young V-339
 Servières, Myriam IV-530

 Shahid, Suleman II-117, II-541
 Shaw, Heather IV-668
 Shaw, Terry III-97
 Shen, Yang Ting V-132
 Sherry, Lance I-193
 Shiba, Haruya II-521
 Shibuya, Yu IV-462, IV-491
 Shiizuka, Hisao I-601
 Shikanai, Nao I-611
 Shimizu, Yuichiro IV-126
 Shin, Yunhee V-362
 Shizuki, Buntarou IV-272, IV-416,
 IV-469, IV-712, V-72
 Shoji, Hiroko V-372
 Siegert, Ingo V-301, V-381
 Siew, Siang-Ting I-470
 Sio, Itiro III-209
 Silveira, Milene S. I-3, III-117
 Simione, Luca IV-637
 Simões, Fabiana I-460
 Singh, Satinder I-193
 Siriwardana, Sanath IV-601
 Sloan, David I-41
 Snyder, Michele II-309
 Soares, Evandro II-428
 Sodnik, Jaka II-551
 Soga, Masato II-134
 Son, Jae Pyo II-631
 Song, Bifeng IV-540
 Song, Sutao V-491
 Song, Yeong-Tae II-56
 Sotero, Gabriel I-460
 Soui, Makram V-141
 Sousa, João IV-678
 Souza Filho, Guido II-197
 Sowmya, Arcot II-631
 Spiegel, Götz III-396
 Sridharan, Seshadri IV-107
 Srivastava, Sumit IV-497
 Stary, Christian I-330, III-456
 Steinhoff, Camie I-265
 Sterbini, Andrea II-351
 Still, Jeremiah D. I-265
 Stillwater, Tai II-640
 Stocklów, Carsten IV-147
 Storz, Michael V-196
 Sturm, Christian III-352
 Sugano, Shohei I-620
 Sugiyama, Masayuki IV-456
 Sun, Ming IV-107

- Sung, Tae-Eung V-362
 Suzuki, Ayaka IV-469
 Suzuki, Ryo IV-348

 Tabak, Feride S. II-391
 Tajima, Terumasa III-466
 Takahashi, Katsumi I-77, I-119, I-137,
 I-500, I-531
 Takahashi, Nobumichi II-647
 Takahashi, Shin IV-730
 Takano, Kosuke II-493
 Takao, Shizuka III-174
 Takemura, Noriko V-270
 Takeno, Hidetoshi I-159
 Takimoto, Munehiro IV-235
 Tamada, Takashi IV-722
 Tamborello, Franklin P. IV-449
 Tamura, Kazuki IV-740
 Tan, Jun-Wen V-474
 Tanaka, Jiro IV-196, IV-272, IV-281,
 IV-416, IV-469, IV-712, IV-730, V-72
 Tang, Honglin IV-243
 Tanikawa, Kyohei III-314
 Tano, Shun'ichi I-480
 Tauchi, Hikaru II-398
 Tavares, Tatiana A. II-197
 Tawatsuji, Yoshimasa IV-398
 Teixeira, Cesar Augusto Camillo III-227
 Teixeira-Faria, Pedro M. I-441, I-490
 Temperini, Marco II-351
 Theng, Yin-Leng II-181
 Thiel, Simon III-29
 Thome, Rainer III-411
 Tian, Yun V-280
 Ting, Chih-Hung IV-506
 Togawa, Satoshi II-503
 Tokdemir, Gul I-310
 Tokumaru, Masataka V-319
 Tomimatsu, Kiyoshi IV-568, IV-578
 Tomiyama, Ken I-620
 Tong, Xin II-207
 Torenvliet, Gerard I-109
 Tornero, Josep I-231
 Tortora, Genoveffa III-342
 Tran, Minh-Triet IV-67, IV-513
 Traue, Harald C. V-474
 Traum, David II-266
 Tripathi, Ramesh Chandra IV-497
 True, Nicholas I-128
 Truong, Chau Thai IV-513

 Tsai, Shih-Lung III-78
 Tsai, Teng-Yao IV-506
 Tsai, Tsai-Hsuan III-359
 Tsai, Yi-Chien III-199
 Tseng, Steven I-540
 Tsianos, Nikos II-418
 Tsukada, Koji III-209
 Tsumori, Shin'ichi II-511
 Tsunoda, Koichi II-398
 Tsuruda, Yu V-391
 Tung, Fang-Wu III-237
 Tung, Tony IV-408
 Tzemis, Evangelos IV-358

 Uang, Shiaw-Tsyr III-475
 Ueda, Kazutaka II-398
 Ueda, Yoshihiro I-77, I-119, I-137,
 I-500, I-531
 Uetsugi, Raku IV-312
 Ushio, Shuhei II-647

 Valla, Massimo V-122
 Valverde, Lauralee II-408
 Vedhara, Kavita II-36, II-189
 Viard-Gaudin, Christian IV-77
 Vilimek, Roman II-578, II-621
 Viller, Stephen III-484
 Vitiello, Giuliana III-342

 Wachsmuth, Ipke IV-117, V-102
 Wahab, Norshahriah IV-523
 Wallach, Dieter I-451
 Walter, Steffen V-301, V-474
 Wang, Hongmei IV-135
 Wang, Jan-Li III-193
 Wang, Liang V-206
 Wang, Lingyun (Max) I-350
 Wang, Pengwei V-254
 Wang, Qian V-216
 Wang, Weizhong V-280
 Wang, Wenzhong V-227
 Wang, Yao-lien V-483
 Wang, Yijun V-448
 Wang, Yixue V-153
 Wang, Yuping V-246
 Wang, Zhaoqi V-227
 Wästlund, Erik III-10
 Watanabe, Keita V-112
 Watanabe, Shintaro V-178
 Webb, Erika Noll II-316

- Weber, Gerhard IV-479
 Weber, Julian II-621
 Wee, William IV-291
 Wei, Chun-Shu V-448
 Wei, Ming-Hsuan II-654
 Weigang, Li IV-648
 Wen, Chao-Hua III-78
 Wendemuth, Andreas V-301, V-381
 Wendler, Stefan I-510
 Wenzel, Christopher II-46
 Whang, Mincheol V-346, V-441
 Wilcock, Graham IV-32
 Winckler, Marco I-51, I-521
 Winschiers-Theophilus, Heike III-332
 Wolff, Christian I-176
 Wolff, Marion IV-429
 Wong, Chui Yin IV-750
 Wong, Man Leong IV-750
 Woodjack, Justin II-640
 Worpenberg, Annika II-323
 Wu, Guanshang II-207
 Wu, Qiong II-207
 Wu, Siju IV-530
 Wu, Tong-Ying II-20
 Wu, Zhanwei I-555
 Wu, Zhongke V-280

 Xiaohui, Tan V-289
 Xiong, Chenlin V-254
 Xu, Bingxin V-236
 Xu, Lele V-498
 Xuesong, Wang V-289

 Yabe, Takao II-398
 Yajima, Hiroshi II-216
 Yalvac, Bugrahan II-408
 Yamaguchi, Takehiko IV-561
 Yamaguchi, Takumi II-521
 Yamane, Shohei III-307
 Yamaoka, Toshiki I-361
 Yamashiro, Mitsuo II-474
 Yamashita, Kuniko V-353
 Yamazaki, Kazuhiko I-77, I-119, I-137,
 I-500, I-531

 Yan, Fu V-289
 Yanagida, Koji I-77, I-119, I-137, I-500,
 I-531
 Yang, Chao-Yang II-20
 Yang, Jason III-484
 Yang, Lei V-216
 Yang, Weixin V-254
 Yang, Yifang V-246
 Yang-Mao, Shys-Fan V-483
 Yao, Li V-491, V-498
 Yasu, Hiroko I-146
 Yasumura, Michiaki V-112
 Yazdifar, Mahshid II-76, V-262
 Yazdifar, Mohammadreza V-262
 Yeo, Alvin W. I-470
 Yildiz, Alparslan V-270
 Yildiz, Ekrem I-310
 Yin, Yabo V-280
 Yoon, Joonsung I-32
 Yoshida, Masanobu II-521
 Yoshikawa, Takuto IV-416, IV-469
 Yoshikawa, Tomohiro IV-740
 Yoshimura, Hiroki V-391
 Yu, Der-Jang I-540
 Yu, Jiamin I-555
 Yue, Yong V-19
 Yun, Myung Hwan IV-368

 Zaki, Michael I-340
 Zaman, Tariq I-470
 Zeng, Wen-Jun IV-506, V-483
 Zetali, Karwan II-464
 Zhan, Yu V-491
 Zhang, Hang V-498
 Zhang, Jiakai V-491
 Zhang, Rushao V-498
 Zhang, Xin V-254
 Zhang, Yugang IV-540
 Zhao, Fan II-46, III-492
 Zhaolao, Lu V-289
 Zhong, Zhengyang V-254
 Zhou, Mingquan V-280, V-289
 Zhou, Xuefu IV-291
 Zieffle, Martina III-49